Fault Tolerant Characteristics of Artificial Neural Network Electronic Hardware

by

Frank C. Zee

A Thesis Presented to the

FACULTY OF THE SCHOOL OF ENGINEERING

UNIVERSITY OF SOUTHERN CALIFORNIA

In Partial Fulfillment of the

Requirements for the Degree

MASTER OF SCIENCE IN BIOMEDICAL ENGINEERING

May 1995

Acknowledgments

I would like to thank Lt. Col. Michael Obal, U.S.A.F., Materials & Structures Program Manager of the Ballistic Missile Defense Organization (BMDO) for sponsoring this research. In addition, I would also like to acknowledge the Jet Propulsion Laboratory (JPL), the California Institute of Technology, and the National Aeronautics and Space Administration for providing the necessary facilities and equipment to conduct the experiments.

I am indebted to Harry Langenbacher, Dr. Taher Daud, and Dr. Anilkumar Thakoor of the Concurrent Processing Devices group at JPL for their assistance with the experiments, their contributions in the analysis, and all their helpful suggestions and comments. I am also very appreciative and grateful for their time and energy in reviewing this paper.

I would like to express my appreciation to my advisor and the committee members, Dr. Theodore Berger, Dr. Vasilis Marmarelis, and Dr. Bing Sheu, for their guidance throughout my years at the University of Southern California.

Finally, I would like to thank my parents, Elaine and William Zee, for their continuing support and encouragement. They made it all possible.

‘l’able of‘ Contents

Chapter Page

List of Figures

Results of Unbiased Chip

Results of Unbiased Chip under Electron Radiation

Figure Page

Results of Biased Chip

Figure Page

Results of Biased Chip under Electron Radiation

Results of Biased Chip under Gamma Radiation at Low Dose Rate

Abstract

The fault tolerant characteristics of analog-VLSI artificial neural network (with 32 neurons and 532 synapses) chips are studied by exposing them to high energy electrons, high energy protons, ” and gamma ionizing radiations under biased and unbiased conditions. The biased chips became nonfunctional after receiving a cumulative dose of less than 20 krads, while the unbiased chips only started to show degradation with a cumulative dose of over 100 krads. As the total radiation dose increased, all the components demonstrated graceful degradation. The analog sigmoidal function of the neuron became steeper (increase in gain), current leakage from tile synapses progressively shifted the sigmoidal curve, and the digital memory of the synapses and the memory addressing circuits began to gradually fail. From these radiation experiments, we can learn how to modify certain designs of the neural network electronic hardware without using radiation- hardening techniques to increase its reliability and fault tolerance.

Chapter 1

Introduction

Significant progress has been made in the research and development of artificial neural networks within the past few years. Inspired by biological" systems, artificial neural networks are highly parallel data processing circuits and are particularly suited to "learn" ill-defined or fuzzy input-output relationships and to perform adaptive interpolations [1].

With the recent technological advances, popularity of artificial neural networks has grown rapidly, and it has found widespread applications in a variety of fields. Some of the operations that neural networks can perform include: classification, where an input pattern is passed to the network, and the network produces a representative class as output; pattern matching, where an input pattern is passed to the network, and the network produces the corresponding output pattern; pattern completion where an incomplete pattern is passed to the network, and the network produces an output pattern hat has the missing portions of the input pattern filled in; noise removal, where a noise-corrupted input pattern is presented to the network, and the network removes some (or all) of the noise and produces a cleaner version of the input pattern as output; optimization, where an input pattern representing the initial values for a specific optimization problem is presented to the network, and the network produces a set of variables that represents a solution to the problem; and control, where an input pattern represents the current state of a controller and the desired response for the controller, and the output is the proper command sequence that will create the desired response [2]. Thus, the diversity of applicati ons of artificial neural networks allows them to be applied to problems in engineering control systems, speech recognition, computer vision (e.g. optical character recognition and image processing), financial market analysis, and weather forecasting, to name a few examples [1]. Neural networks are capable of easily performing many tasks that conventional regression techniques and traditional artificial intelligence systems find difficult or impossible to solve.

“1’here has also been a high level of interest in the applicatio n of artificial neural networks to the field of medicine. Image processing neural networks have been used to diagnose hepatic masses [3,4] and breast tumors [5]. Researchers also have trained mural networks to interpret ventilation-perfusion (V/Q) lung scans by exposing it to 100 consecutive V/Q scans with pulmonary angiographic correlation [6]. It was then used to classify 28 new scans without access to the angiograph ic correlation. When the resultant classifications were compared with the rankings of an experienced observer who also read the scans without knowledge of the correlative angiographic data, the network significantly outperformed the experienced observer in the prediction of the likelihood of a pulmonary embolism.

Some researchers have begun to realize that the clinical in formation generated in the high dependency environment (intensive care unit, operating room, emergency room, recovery room) can actually be interpreted as a "pattern," with each unique combination of symptoms, signs, and laboratory results representing a different clinical scenario.

One such researcher has created a neural network to diagnose myocardial infarction in persons reporting to the emergency room complaining of anterior chest pain 17,81. A neural network was trained to i nterpret history, physical examination findings, and specific emergency room electrocardiogram (f :CG) findings and to predict the probability that the patient had suffered a myocardial in farction. The diagnostic performance of the artificial neural network was compared with the opinion of the attending emergency room physician. The physician diagnosed myocardial infarction with a sensitivity and specificity of 77% and 84%, respectively, whereas the neural network performed with a sensitivity and specificity of 97.2% and 96.2%. This is a result of the neural network learning from examples. Subsequent examination of the trained network revealed that it predicted myocardial infarction by placing (diagnostic importance on clinical variables that had not be en shown previously to be highly predictive of infarction.

These examples show that artificial neural networks have the ability to support medical decisions. Thus far, the types of network architectures and learning algorithms used by biomedical researchers have been relativel y simple. '1'0 extend this technology to greater and more compl icated medical applications, future research needs to investigate the properties and behavior of complex, state of the art, application-specific network topologies.

The potential benefits of neural nets extend beyond the high computation rates provided by massive parallelism. Neural nets typically provide a greater degree of robustness or fault tolerance than von Neumann digital sequential computers since there arc many more processing nodes, each with primarily local connections, and the information is CO(1C(1 distributively on weighted synaptic connections or links, Damage to a fcw nodes or links thus need not impair overall performance significantly.

Most neural network learning algorithms also adapt their synaptic connection weights in time to improve performance based on the current results. This is a major focus of neural network research, and the ability to adapt and cent inuc learning is essential. Adaptation also provides a degree of robustness by compensating for minor variations in characteristics of neurons. Traditional statistical techniques are not adaptive, typically processing all training data simultaneously before being given new data. Neural network classifiers are also non-pai ametric and make weaker assumptions concerning the shapes of underlying distributions than traditional statistical classifiers. They may thus prove to be more robust when distributions arc generated by nonlinear processes and are strongly non-Gaussian.

Recent advances in neural networks have increased the level of interest for critical applications, such as deployment in space, for military applications, in high energy physics laboratories, in chemical processing industries, or in hospitals where reliable performance and longevity is a requirement.

The purpose of this research is to demonstrate the fault tolerant characteristics of analog neural network VI .S1 chips. Exposure to ionizing radiation was performed in order to study the behavior and graceful degradat ion of the electronic hardware.

in the next chapter, before introducing the artificial neural network chip, we give a brief definition and background on the artificial neural network model and its derivation from the biological model. Chapter 3 will describe the experiments conducted on the electronic hardware, including the three types of ionizing radiation sources, chip setup, and the tests to be performed to study the degradation on various components of the chips. The results of the radiation experiments are given in Chapter 4, providing a detailed (inscription of the effects of radiation at different cumulative doses on the performance of the chip. The results are compiled and analyzed in Chapter 5. We elaborate on the mechanisms responsible for different degradation effects. Chapter 6 provides a conclusion drawn from our experiments and a direction charted towards future research.

Chapter 2

Neural Networks

Although there are numerous ways to define artificial neural networks, Kohonen [9] has attempted to state it with the following general definition:

> *"The artificial neural networks" are massively parallel interconnected networks of simple (usually adaptive) elements and their hierarchical organizations which are intended to interact with the object of the real world in the same way as the biological nervous systems do.*

The structure of artificial neural networks are modeled after the organization of the human nervous system, specifically the human brain. They are composed of elements that perform in a manner that is analogous to the most elementary functions of the biological neuron. Although the resemblance is superficial, artificial neural net works as physical cellular networks exhibit such brainlike characteristics as their ability to learn from experience, generalize on their knowledge, perform abstraction, and make errors, which are all more characteristic of animal behavior than that of conventional digital computers.

T() introduce the neural network hardware, wc first explore in a simplistic way the biological equivalent before tracing artificial neural networks and their electronic embodiment from the "connectionist" point of view rather than biological.

2.1 Biological Neural Networks

A human brain contains over one hundred billion computing elements called neurons [10]. Exceeding the number of stars in our Milky Way galaxy, these neurons are the fundamental building blocks of the biological neural network, the nervous system. A neuron is an elementary nerve cell which typically has three major regions: the cell body (or soma), the axon, and the dendrites.

The cell body, similar to any other cell, conducts maintenance activities for the neuron. However, the outer membrane of the neuron's cell body also has the unique capability of generating nerve impulses called action potentials. The axon is a long cylindrical fiber that serves as a transmission line to carry the impulses from the cell body. The end part of an axon splits into a fine arborization where each branch terminates in a small end bulb almost touching the dendrites of the neighboring neurons. Dendrites form a dendritic tree, a very fine bush of thin fibers emanating from the cell body. Functionally, the dendrites receive information from the axons of neighboring neurons.

This axon-dendrite contact organ is called a synapse and is where the neuron introduces its signal to the neighboring neuron. There are perhaps one hundred trillion synapses forming the interconnections within the biological neural network [10]. The signals reaching a synapse and received by dendrites are electrical impulses. But, the interneuronal transmission are usually affected by the release of chemical transmitters at the synapse, where the axon terminals generate the chemical that affects the receiving neuron's dendrite. Each neuron is connected to thousands of other neurons in this way.

From there, the signals are passed on to the cell body of the receiving neuron which either generates an impulse to its axon to be passed on to succeeding neurons, or produces no response. The neuron's response is generated if the total potential of its membrane reaches above a certain till'CSllol(l level. The membrane can be considered as a shell, which aggregates the magnitude Of the incoming signals over a short time interval called the period of latent summation. incoming signals can be excitatory if they cause the firing, or inhibitory if they hinder the firing of the response. Thus, the neuron generates a pulse response and sends it to its axon only if the excitation exceeds the inhibition by the amount called the threshold of the neuron.

Since a synaptic connection causes the excitatory or inhibitory reactions of the receiving neuron, it is practical to assign positive and negative unity weight values respectively, to such connections. This allows us to reformulate the neuron's firing condition such that the neuron fires when the total sum of the weights to receive impulses exceeds the threshold value during the latent summation period.

The incoming impulses to a neuron can only be generated by neighboring neurons and by the neuron itself (feedback), Usually, a certain number of incoming impulses are required to make a target cell fire. Impulses which are closely spaced in time and arrive synchronously are more likely to cause the neuron to fire. As mentioned before, observations have been made that biological networks perform temporal integration and summation of incoming signals. The resulting spatio-tempera] processing performed by natural neural networks is a complex process and much less structured than digital computation. The neural impulses are not synchronized in time as opposed to the synchronous discipline of digital computation.

The characteristic feature o f the biological neuron is that the signals generated are either absent or have maximum values. This means that information is transmitted between the nerve cells by means of binary signals.

Despite its apparent simplicity, this computational function accounts for most of the known activity of the brain. Underlying it, however, is a complex electrochemical system. This network of neurons is responsible for all of the phenomena that we call thought, emotion, and cognition, as well as foı performing myriad sensorimotor and autonomic functions.

The brain is somehow capable of taking neurons which are five or six orders of magnitude slower than silicon logic gates, and organizing them so as to perform some computations many times faster than the fastest digital computer. One way the brain seems to have managed to do this is by massive parallelism in its gray matter. That is, the computing elements are arranged so that very many of them are working on a problem at the same time. Since there are huge numbers of neurons, the weak computing powers of these many slow elements are combined together to form a powerful resultant. The architecture of the brain, however, is not well understood at this time. The speed of the neurons have not increased much involution, so the way to get more power seems to be to add more neurons, a strategy highly developed in our own massive cerebral cortex [1()].

The above discussion is extremely simplified when seen from a neurobiological point of view, though it is valuable to gain insight into the principles of "biological computation. " Artificial neural networks arc far simpler than their biological counterparts.

2.2 Artificial Neural Networks

A variety of artificial neural network architectures and algorithms have been reported in the literature [1]. In general, the arch itectu e can be defined as an interconnection of neurons such that neuron outputs arc connected, through synaptic weights, to all other neurons including themselves.

Every neuron model consists of a processing clement with synaptic input connections and a single output. The signal flow of neuron inputs and output are considered to be unidirectional. The neuron as a processing node performs the operation of summation of all its weighted inputs (representing synaptic strength]), and subsequently, performs a nonlinear operation through its activation function. The summed value, which determines the activation level of the neuron, passed into the function may be considered as an analog to the biological" neuron's membrane potential. Different artificial neural network algorithms make usc of different (definitions of the activation function. Some examples arc the hard-limiting activation functions (binary functions) called threshold logic units and the soft-limiting activation functions (continuous functions) called sigmoidal characteristics. Most artificial neural systems do not involve the biological neuron features of' delay, refractory period, or discrete-time operation. In fact, the neuron model represent instantaneous, memoryless networks, since they generate the output response determined only by the present excitation.

Two particularly popular networks are the feedforward and the feedback network architectures. In the former, all external inputs arc fed to a layer (layer #l) of neurons through synaptic weights in such a way that each input is fed to all of the neurons, Similarly each neuron in layer 1 is connected to each neuron of the next layer (layer #2) through synaptic weights. Layer #2 may be similarly connected to layer #3 which may be the final layer giving the outputs and thus termed the output layer. The intermediate layers between the inputs and the output layer are termed hidden layers, This architecture is called ii feedforward network because of the forward flow of signals. A feedback network can be obtained from the feedforward network by connecting the neurons' outputs to their inputs.

This simple discussion gives an introduction to the architecture of the artificial neural network and establishes a basis for the electronic hardware design.

2.3 Neural Network Electronic Hardware

The artificial neural network chips were specially designed by the Jet Propulsion Laboratory (JPL) for conducting radiation experiments in space on an orbiting satellite, but were designed and fabricated without any radiation-hardening techniques or any protection from radiation. It requires a single power supply of 8 volts and consumes less than 80 **mw** of power. A brief description of this chip follows, but more extensive coverage is given in the Appendix (page 11 6).

The neural network chip contains 32 neuron cells and 532 synapse cells arranged in a partially populated (due to power limitations for space flight) 32 row by 32 column array. A block diagram of the layout of the artificial neural network chip is shown in Figure 1. The neurons lie along a diagonal in the array. The synapses are placed where the row number is less than the column number. 1 lowever, when the column number is greater than 22, the synapses arc placed in every row position except along the diagonal.

Neuron () (bottom left in Figure 1) can only receive input from an external source. Neuron 1 also can receive input externally or from neuron 0 through the synapse at row (), column 1. Neurons 2 through 31 receive inputs onl y from any of the other neurons through the synapses in their respective columns. The first 22 neurons each can receive input from as many synapses as the column number that they are positioned in (i.e., neuron 20, located in column number 20, can receive input from 20 synapses). The last 9 neurons in columns 23 through 31 all have fully connected feedback circuits and thus have 31 synapses connected to the input of each neuron,

The nature of the neural network chip is primarily analog for compactness and for low power consumption, with the digital portion only playing a supporting role. The major analog components are the neurons, synapses, and an output buffer. The major digital components are the static memories for storage of synaptic weight within each of the 532 synapses.

2.3.1 Neu ron

The neurons arc nonlinear transimpedance amplifiers. The characteristics of an operational amplifier resembles a sigmoidal function and thus is a natural circuit for a neuron. Each neuron outputs a voltage which is a sigmoidal function of its input current. A block diagram of the neuron is shown in the upper right corner of Figure 2. A negative input current forces the neuron's output high, and a positive input current forces the output low. Section A. 1 of the Appendix gives a more detailed description,

A voltage to current converter is provided for each row to convert the neuron output voltage to current input for the synapses, When the voltage output of the neuron is high, there will be sufficient current input for the synapses in that row to multiply. The higher the neuron output volt age, the more current that will be available for the respective synapse to multiply. On the other hand, if the neuron voltage output is low, there will be virtual ly no input current to the synapses in that row.

2.3.2 Synapse

The synapse circuit consists of a 7-bit multiplying digital to analog converter (MDAC) and a 7-bit digital memory. This circuit o utputs a binary weighted multiple of its analog input current. The input current coming from the neuron in that row is multiplied by the stored digital weight (integer factor between -63 and +63). The digital memory provides programmable weight storage and is randomly accessible. The upper left corner of Figure 2 shows a block diagram of this synapse. A negative weight value programmed into the synapse causes it to have a positive output current while a positive weight value Causes the output to be negative. The circuit design is more thoroughly covered in Section A.2 of the Appendix.

Each synapse's current output is an input to a neuron in its column. Thus, there is a single current summing node in each column to which all the synapses are directly connected, and that node is also connected to the input of the neuron in that column.

2.3.3 Output Buffer

The output buffer is a wi(lc-range transconductance amplifier configured as a unity-gain follower. The output voltage closely follows its input voltage and provides sufficient driving power to interface the neural network chip to other chips. The input to the output buffer is the voltage from one of the 31 neurons selected by an analog multiplexer. A special circuit designed into the output multiplexer allows the output buffer to be tested separately by feeding its input with an externally supplied voltage.

The single output buffer and multiplexer design was utilized to simplify the interface between the chip and its external circuitry by reducing tile number of separate output signals. Section A.3 in the Appendix describes the characteristics and design of the buffer in more detail. Since all of the neural network's output must go through this buffer, it is the most critical component (representing a single point failure mechanism for tile chip).

2.3.4 Operation of Neural Network

The synapses in rows () and 1 can be used to provide analog input values to the neural network (Figure 1). Neurons () and 1 can be biased by an external current input so that tile synapses in their respective rows will output current when they are programmed with a nonzero weight val ue. In this manner, the programmed synapses will force the neurons in their columns to a certain voltage related to the synapse-weight. This is how the input layer of neurons are set to the desired input value.

The input ayer of neurons are connected to a hidden layer of neurons through synapses in the same rows as the input neurons and in the same columns as the hidden neurons. 1 .ikewise, the hidden neurons arc fed through synapses to a layer of output neurons. Finally the output neurons could be read out one at a time through the analog multiplexer of the output buffer. All the unused synapses were programmed to a weight of zero so that they were effectively nonexistent and should have no effect on the operation of the network. A block diagram of the operation of the neural network is shown in Figure 2.

If any neurons were determined to be severely damaged by radiation, other neurons could be used in their place, and the synapses connected to the faulty neurons could be programmed to zero and thus, theoretically have no effect (m the network, Row () was usually used to control the network inputs. However, if row 0 was faulty, then those synapses would be programmed to zero, and the synapses in row 1 would be used as the network inputs,

The synapses in columns 23 through 30, where the row number is greater than the column number, could be used to supply a feedback path (Figure 1). However, the implementation of neural networks using feedback was beyond the scope of this experiment. These "feedback" synapses were only used in the memory tests and in the testing of the individual synapses.

BLOCK DIAGRAM OF ARTIFICIAL NEURAL NETWORK CHIP

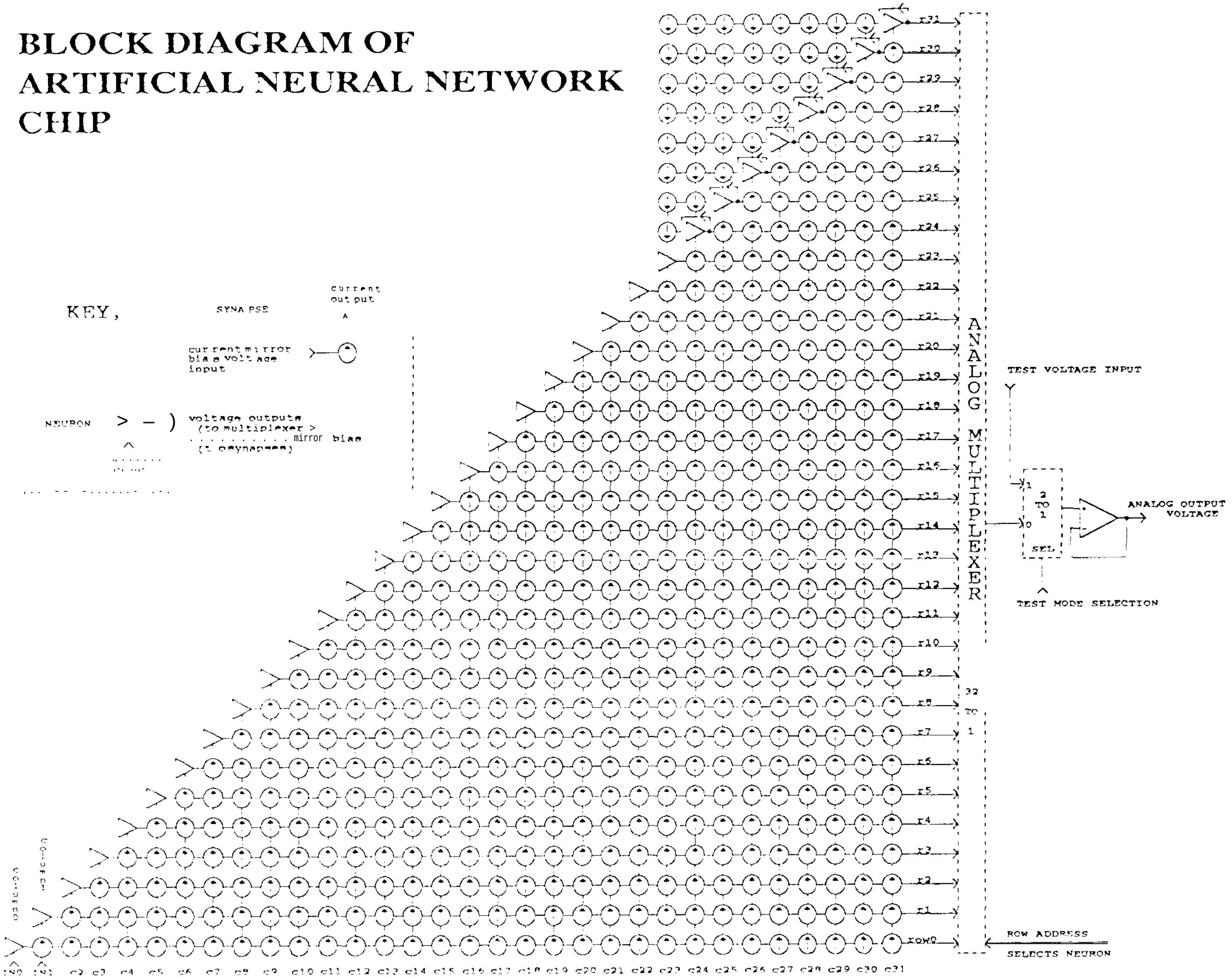

Figure 1 Block Diagram of the Artificial Neural Network Chip

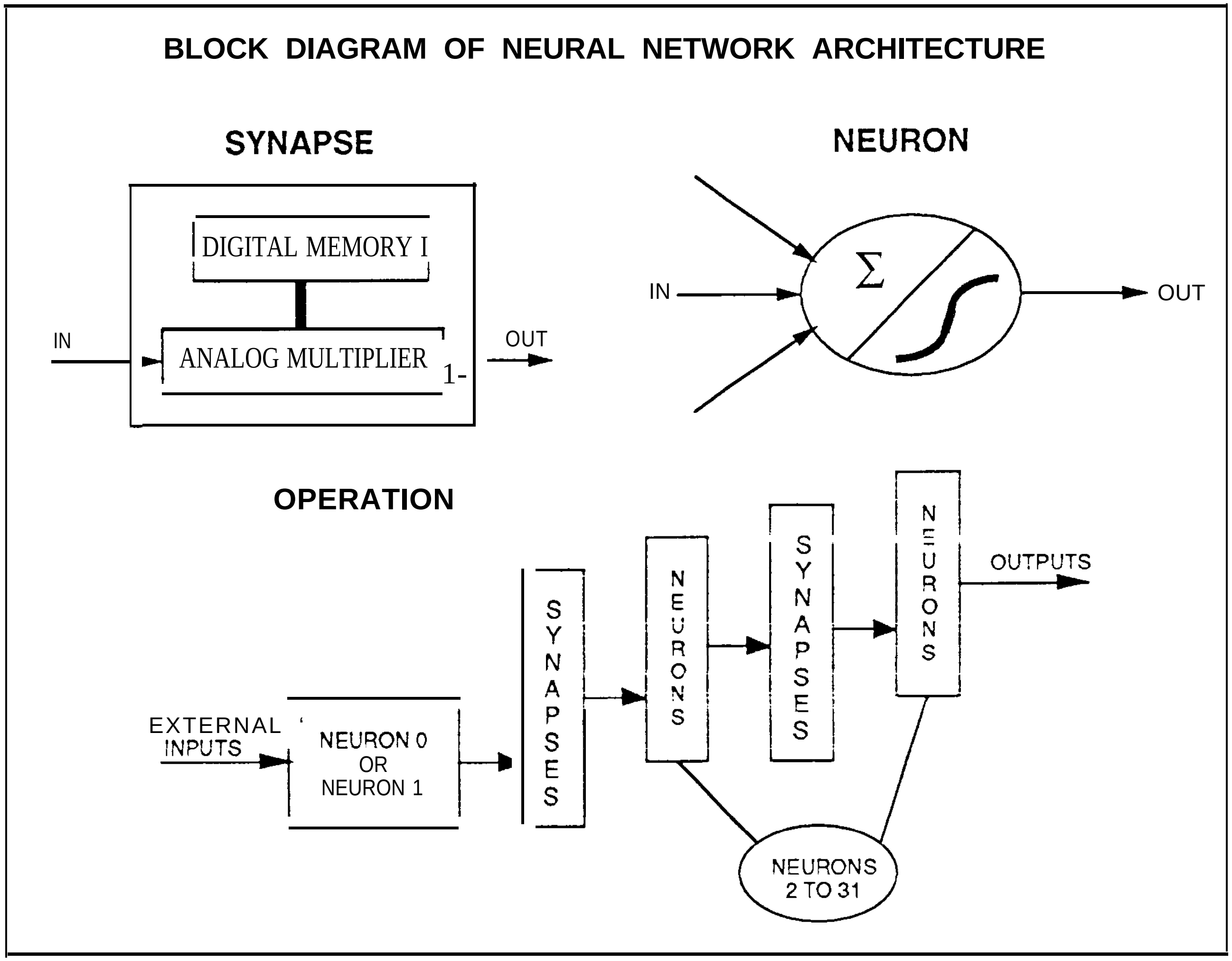

Figure 2 Block Diagram of Neural Network Architecture

Chapter 3

Experiment

To study the fault tolerant characteristics of the artificial neural network chips, they were exposed to three different types of ionizing radiation. A total of 7 chips were used for the experiments where each chip was either setup with or without power applied. Several tests were implemented to study the effects of the radiation on the various components of the hardware. The radiation experiments were concluded when the neural network chips failed and were unit'stable. The following sections will describe these parts of the experiment in more detail.

3.1 Radiation Sources

Ionizing radiation is that which possesses enough energ y to break atomic bonds and create electron/hole pairs (i.e., cause ionizat ion) in the materials of interest, which in the case of MOS devices are primarily silicon dioxide and silicon. This radiation may be in the form of photons with energies greater than the bandgap of the material, or in the form of particles such as electrons and protons. Three types of ionizing radiations were used in the experiments: high energy electron, high energy proton, and gamma radiations. The radiation exposure was conducted at a uniform dose rate and at room temperat ure.

The electron irradiation was conducted at the JPL's Dynamitron Accelerator 1 .aboratory using a 1.() MeV electron source at a flux level of 2.08 x 1(7 e^-/(cm^2-sec) for a dose rate of 50 rad/sec. Two chips were exposed to the electron radiation.

The proton irradiation used an 8.() MeV proton source at a flux level of 1.0 x 10^8 p^+/(cm^z-see), again for a dose rate of 50 rad/sec. This also was conducted using two neural network chips at the California institute of Technology's Tandem Accelerator Laboratory.

Gamma irradiation was conducted at the JPI .'s Total Ionizing Dose Laboratory using a 1.25 MeV 4,()()() curie Co^{60} Gamma Cell. Two different dose rates were used by placing a chip 14 cm from the source for a high dose rate of 50 rad/sec and another chip 335 cm from the source for a low (lose rate of 0.1 rad/sec. This was done to study the effects of high and low dose rates on chip (degradation. A total of three neural network chips were exposed.

For each of these radiation sources, the neural network chips were exposed to the radiation for a specific time period to receive a particular cumulative dose after which the chips were tested. The chips would then be exposed again. This cycle was repeated until the chips showed severe degradation and were untestable. At the completion of each of the radiation exposure experiments, room temperature annealing measurements were made on all of the chips.

3.2 Biased vs. Unbiased Chips

During the radiation exposure, at least one chip was "under bias" and one chip was unbiased. "Under bias" means that power was applied to the chip so that electric fields were present in the devices.

The biased chips had 8 volt power supplied to the Vdd pin. The inputs to the bias transistors, however, were turned "off", and either connected to ground for n-FET devices or to Vdd for p-FET devices. Additionally, all of the output pins were left floating and all of the digital inputs were grounded. The net result was that many of the individual transistors in the chip had no voltage applied to their gates. However, since the synapse memory latches were all uninitialized, the state of a large majority of transistors were randomly determined each time the chip was powered up. I'bus, the transistors in the neurons were also in a random state since their input depended on the synapses.

The goal of this specific biasing scheme was to avoid CMOS latchup. This occurs when some parasitic bipolar transistors are formed from the CMOS transistor components which causes excessive power consumption. Latchup still occurred when there was severe degradation at high cumulative doses of radiation (See Section 5.5 of Chapter 5).

The unbiased chip was left with all the pins connected to a conductive foam pad. This prevented any charge b uild-up on the transistors due to radiation.

One chip was biased and another (me was unbiased in both the electron and proton radiation exposures. In the gamma irradiation, (me chip was biased while a second one was unbiased at the high dose rate (50 ra(i/see), and a third one was biased at the low dose rate (0.1 rad/see).

3.3 Chip Tests

To evaluate the effects of radiation exposure, the neural network chips were tested on a 33 MHz 486 IBM A’]’ compatible computer with digital and analog interface boards plugged into the ISA bus. The interface boards supplied (me 12-bit A/D channel, two 12-bit D/A channels, and all the necessary digital I/O's.

A fixture was constructed to test the chips which consisted of a 64-pin ZIF (z,er(~-illscr(i(~l~-force) socket to hold the chip, connectors for the computer interface, power supply connections, and resistors to bias some chip inputs. in addition, digital level shifters were provided to enable the 5 volt digital signals from the computer to interface with the neural network chip at different voltage levels. Also, the fixture contained a toggle switch to select either neuron 0 or neuron 1 for input from a DAC (digital-to-analog converter) output from the computer.

A test program was written in C language with about 2700 lines of code which focuses on the testing of the major analog and digital components of the neural network chips by performing four tests: neuron, output buffer, synapse-neuron, and memory tests. We describe these tests in further detail.

3.3.1 Neuron Test

In the neuron test, the transimpedance transfer function of the neuron was measured. Only neuron 0 and neuron 1 were tested, since they were the only neurons that had external inputs. A voltage from a computer controlled DAC was connected to the neuron input through a resistor. When this voltage was ramped, the neuron was swept with a current input in the range of -400" mA to +400 mA. The output voltage was measured for each of the 800" DAC voltage steps and ranged from () to 8 volts. The resulting curve should represent a sigmoidal function centered at a current input of 0 mA and an output voltage of 4 volts (Vdd /2). Since the neuron output voltage could only be measured through the output buffer, the transfer functions of the neuron and the output buffer were convolu ted to produce the re sult.

3.3.2 Output Buffer Test

The output buffer test checked whether the buffer's output voltage followed the input voltage. The multip lexer that selects the input to the output buffer was switched over to a pin on the chip such that an external voltage could be applied. Then, the output voltage was measured while sweeping the input voltage from 0 to 8 volts. The output should increase linearly with the input voltage. Since all of the neural network's outputs must pass through this buffer, it is the most critical component, representing a single point failure mechanism for the chip.

3.3.3 syllapsc-Neuron Test

The synapse-neuron test examined all the 532 synapses and 31 of the 32 neurons. Each test involved one synapse-neuron pair where the output current of each of the synapses was swept by programming them from -63 to +63 and measuring the voltage output from the corresponding neuron. Inputs to all the synapses originated from neuron zero, either directly as for the synapses on row zero or indirectly through a synapse-neuron pair. When a synapse in a particular row is being tested, the neuron in that row must have a sufficiently high output voltage to supply current into the synapse. '1'bus, the corresponding synapse in row () must provide an appropriate input current to that neuron to force its output high. For synapses in row 0, neuron 0 is forced high by its external current input. All the other unused synapses were programmed to zero to eliminate their effects. The results of this synapse-neuron test were analyzed in two ways.

First, all the voltage responses from the neurons were graphically displayed, which allowed us to visually inspect the degradation of all of the 532 synapses. Without radiation, all the curves were sigmoidal in shape (as in the neuron test) and centered at a synapse DAC value of zero. Statistical variations in fabrication were evident as all the 532 synapses showed slightly varied sigmoidal curves. The synapses in row 0 also showed offset curves since neuron 0 was externally biased.

Second, to quantify our observations, a "monotonicity" lest was devised. In this test, each point on the synapse sigmoidal curve was compared with the next more positive point to see if the neuron output voltage was monotonically increasing. Errors were counted as any pair of points that were not monotonically increasing. All the chips initially had no monotonicity errors.

The synapses were controlled through other neurons at the input and convoluted through other circuits at the output, and therefore, were the hardest components to test. These curves showed the radiation effects not only on the synapse but on the neurons, the gain-contr ol circuits, and the output buffer.

3.3.4 Memory test

The memory test was performed to evaluate the digital memory in all the 532 synapses, where each synapse memory was organized as a randomly accessible 7-bit word with a lo-bit address. The memory lest consisted of two tests: a standard "all values" memory test which verified if correct values were written to the chosen memory locations; and a walking-ones memory test which verified if the correct memory locations were being addressed.

The standard "all values" memory test wrote and read all 128 values (-63... -0,+ 0... +63) for each of the 532 synaptic memory words with some attempt to test memory addressing functionality as well. Errors were counted as a mismatch between the value written and the value read from each of the synapses.

In the walking-ones memory test, a background pattern (e.g. 0101010)" was written to all of the synaptic memory addresses, while a foreground pattern which was the complement of the background pattern was written to onc address. All the addresses were then read back to detect if any of the patterns were written to an inappropriate memory address. The foreground pattern was written to all of the synapse memories in succession. This exhaustively tested for errors in the memory addressing circuits,

Although there arc 492 unused memory addresses (out of the 32x32 matrix), both of the memory tests still included them in the testing to confirm that they had no effect on the real memory locat i ons. The total number of errors and the total errors by row and column were recorded. in addition, if and when errors developed for either of the tests, the first 200" erroneous results were recorded so that the errors in the individual bits of the word could be analyzed.

3.4 Failure Modes

Since the chips could fail in a number of different ways, wc had to develop a criteria of functionality and failure for the chip as a whole. The chip was considered to be nonfunctional if any of the following occurred: tile chip consumed over 200" mA with an 8 voit power supply (1.6 Watts); the analog performance was so distorted (sigmoidal curves from the neuron and synapse- neuron tests) as to make it unusable as a neural network; or 80% of the total memory had failed. in most of the experiments, the chips had latched-up during exposure or during testing, exceeding the power limit, and were untestable and hence. declared nonfunctional.

chapter 4

Results

The most significant observation from all of the radiation experiments was that the biased chips degraded at least an order of magnitude faster than the unbiased ones. The cumulative dose for the unbiased chips were 140 krads for the electron radiation, 250 krads for the gamma radiation , and 440 krads for the proton radiation. The biased chips had lower cumulative doses of 6 krads for the electron radiation, 7.5 krads for the low dose rate (0.1 rad/sec) gamma radiation, 5 krads for the high dose rate (50 rad/sec) gamma radiation, and 30 krads for proton radiation. We shall look at the degradation of the neural network electronic hardware in more detail for each of the three types of radiation, first for the unbiased chips and then for the biased chips.

4.1 Unbiased Chips

The electron radiation will be discussed first, followed by the gamma and proton radiations. Since similar degradation characteristics were observed for all three radiation sources, only figures, showing the effects of radiation on synapse-neurons, neurons () and 1, the output buffer, monotonicity errors, and memory errors, for the electron radiation are included as an example.

Figures 3.1a and 3.1 b show for the three radiation exposures the percent monotonicity errors versus total radiation dose and annealing time, respectively. In Figures 3.2a and 3.2b, the percent memory errors from the standard memory test are shown again for the total radiation dose and annealing time. The results of the waiking-ones memory test arc shown in Figures 3.3a and 3.3b. These figures not only allow us to see the hardware degradation as will be discussed in the next sections, but also to compare the radiation effects from the different sources as will be described in tile next chapter.

4.1.1 Electron Radiation

With electron radiation, the unbiased chip had no noticeable sign of any radiation effects up to 20 krads. Figure 4. la shows the 532 synapse-neuron baseline characteristics at () krads.

The output buffer test at 30 krads showed initial signs of a (distortion ("kink") at the high end where the input voltage was about 7 volts and the output voltage was 6.75 volts. Just beyond that point, the output voltage suddenly increased more rapidly than the input voltage. Figure 4.4 shows this "kink" in the output buffer characteristics. This could be seen in all the tests including the neuron tests and the synapse-neuron test, since the output voltage was always measured via (he output buffer.

The neurons 0 and 1, and the synapse-1 ieuron curves all began to show slightly steeper (higher gain) sigmoidal curves with a 1% monotonicity error at 40 krads. This can be observed in Figure 4.2 and 4.3, showing the neuron () anti 1 characteristics, respectively, and in Figure 4.1b, showing the 532 synapse-neuron characteristics at 40 krads. in addition, the "kink" from the output buffer grew larger and was definitely observable in al] of the curves.

At 50 krads, all the sigmoidal curves showed noticeable steeper slopes, and in addition, showed an 11% monotonicity error (Figure 3.1a). Some of the sigmoidal curve sat 60 krads of total dose in the synapse-neuron plot (Figure 4.1c) started to shift to the left with a 30% monotonicity error. Instead of having the sigmoidal curve centered at the input synaptic weight of (), they were centered at a negative synaptic weight value. This means that the neurons were fermi high at earlier (lower than normal) input synaptic weight value.

As the cumulative dose was increased to 70 krads and then to 80 krads, increasing shifts to the left were noticed (neurons turning on earlier) as shown in Figure 4.1d. Monotonicity errors similarly increased from 48% at 70 krads to 56% at 80 krads (Figure 3.1a). Some curves were shifted to such an extreme extent that the neurons were always forced high, showing a maximum output voltage of 8 volts (Vdd), resulting in a flat line at 8 volts, instead ofa sigmoidal shaped curve.

At 100 krads, in addition to the flatness in the synapse-neuron curves (Figure 4.1e), the monotonicity errors increased to **64%.** At this point, the radiation experiment was temporarily terminated and the chip was left to anneal at room temperature overnight. After about 13 hours of annealing, a few of the synapses and neurons had partially recovered. The sigmoidal curves were shifting back to the right and reducing the monotonicity errors back to 58% (Figure 3.1a). however, there was practically no change in the slope of the neuron and synapse-neuron transfer curves and the "kink" at the high end for the output buffer.

Further, total dose increase to 135 krads resulted in similar but more damage. Monotonicity errors increased from 61% at 110 krads to 72% at 120 krads, at which point they remained saturated (flattened out) up to 135 krads as shown in Figure 3.1a. in Figure 4.1f, the 532 synapse-neuron characteristics at 120 krads are shown. Memory errors started to develop with 11% at 120 krads and reaching 27% at 135 krads for the standard memory test (Figure 3.2a). For the walking-ones memory test, there were 7% errors at 120 krads, which increased graduall y to 48% at 135 krads (Figure 3.3a).

At the final dose of 140 krads, more than 75% of the synapse-neuron characteristic curves (Figure 4.1 g) were flat (neurons fully turned on). The monotonicity errors reached a maximum of 74% where the standard and walking-ones memory tests reached 29% and 48%, respectively.

Column dependent errors were observed for both the monotonicity and the memory tests. The monotonicity errors showed only a slight column dependency (Figure **4.5) where** the last 9 columns had 80% errors and the other columns had about 65% errors at 140 krads. The memory errors were mainly limited to the last 9 columns, reaching 50% errors for the standard test and 90% errors for the walking-ones test at 140 krads. Figure 4.6 shows the memory errors by column for the standard memory lest. Synapse memory bit 6, which is the sign bit, showed significant errors for both memory tests for the first 200" errors detected. This can be seen in Figure 4.7.

Room temperature annealing was done and measurements were taken 5 hours, 18 hours, 31 hours, 60 hours, and 106 hours after the final radiation dose. After 60 hours of annealing, the walking-ones memory errors fully recovered (Figure 3.3 b), but the standard memory test still showed 6% errors, but recovered fully after 106 hours of annealing (Figure 3.2 b). Monotonicity errors recovered only slightly to 63% after 106 hours of annealing (Figure 3.1 b). The synapse-neuron curves shifted back to the right during room temperature annealing measurements. These can be seen in Figures 4.1h, 4.1i, and 4.1j for annealing measurements at 31 hours, 60 hours, and 106 hours, respectively. '1'here were no changes to the curves for neurons () and 1. The sigmoidal curves continued to be very steep and showed high gain. The output buffer also did not recover during the annealing and the "kink" at the high end did not change.

4.1.2 Gamma Radiat ion

The unbiased chip for the gamma radiation performed without any degradation at 20 krads, but started showing some effects at 30 kinds where sigmoidal curves of a few synapse-neurons began to turn steeper.

At 40 krads, further steepness was observed and the "kink" in the output buffer started to develop. '1'here was 1 % monotonicity error as shown in Figure 3.1a. As the dose was increased to 50 krads, more steepness in the sigmoidal curves was seen with 2% monotonicity errors, and the output buffer showed a larger "kink" at the high end. Both the neuron () and 1 curves were also showing steeper sigmoidal characteristics and shifting to the left.

At 60 krads, all of the synapse-neuron curves became very steep. These curves started to shift to the left at 70 krads, and monotonicity errors increased to 1()%.

As the dose was increased to 80 krads, 90 krads, and 1()() krads, all the synapse-neuron curves shifted further to the left, started losing their sigmoidal nature, and became fiat at 8 volts (neurons always high). The monotonicity test showed 25% errors at 80 krads, 37% at 90 krads, and 42% at 1 ()() krads (Figure 3.1a). The chip was then left overnight to anneal at room temperature. After over 14 hours of annealing, there was only slight recovery as a few synapse-neuron curves shifted towards the right anti the monotonicity errors decreased to 34%.

When the total dose was increased to 120 krads, 140 krads, and 160 krads, the left shift of the synapse-neuron curves continued and the characteristics became flat. Monotonicity errors increased from 49% at 120 krads to 52% at 140 krads and then seemed to saturate to 60% at doses greater than 160 krads. At 170 krads, about half of the synapse-neumn curves were fiat.

From 180 krads to the final cumulative dose of 250 krads, the remaining synapse-ncurm curves shifted to the left. Only a few good synapse-neuron curves were seen, and the remaining curves were all flat at 8 volts. The monotonicit y errors reached 63% at 250 krads. Memory errors started occurring at 190 krads with 8% errors for both the standard (Figure 3.2a) and walking-ones (Figure **3.3a)** tests. The memory errors, similar to the monotonicity errors appeared to increase exponentially and seemed to saturate at 30% with 250 krads of total dose for the standard test and at 48% with 220 krads for the walking-mm test. For both the memory tests, there were significant errors with memory bit 6 (sign bit) in the first 200" reported errors.

Column dependency was observed for both the memory and monotonicity tests. Again, there were memory errors only in the last 9 columns for both memory tests with 50% errors for the standard test and 90% errors for the walking-ones test at 250 kinds. Monotonicity errors showed only slight column dependency because at 250 krads the last 9 columns had 68% errors and the other columns had 55% errors.

Annealing was done at room temperature following the final cumulative close of 250 krads and the characteristics were measured respectively after 17, 28, 44, 52, 75, and 115 hours of annealing. After 17 hours, there were no errors resulting from the walking-ones memory test as shown in Figure 3.3b. The standard memory test also recovered after 28 hours of annealing (Figure 3.2 b). Monotonicity errors recovered only slightly to 49% after 115 hours of annealing (Figure 3.1 b). The synapse-neuron curves showed that some sigmoidal curves had partly recovered by shifting back to the right. The steepness or gain of the sigmoidal curves did not change for any of the neurons. Similarly the "kink" in the output buffer was unchanged.

4.1.3 Proton Radiation

For the proton radiation, the unbiased chip had cumulative doses of 5, 10, **20, 80,** 11 (), **150, 200,"280,** and **440** krads. The chip performed without any noticeable degradation or radiation effects up to 20 krads.

At the next level of total dose (80 krads) measurement, the output buffer showed a "kink" which was observable in all the curves. Also, the sigmoidal curves for all the neurons became steeper, and there were **2%** monotonicit y errors (Figure 3.1a).

The curve steepness further increased at 110 krads with 23% monotonicity errors. The left shift of the synapse-neuron curves started at 150 krads causing monotonicity errors to increase to 58%.

At 200" krads, most of the synapse-neuı on curves were flat at 8 volts (being shifted to the extreme left and having the neurons fully turned on). There were 86% monotonicity errors. Memory errors also developed. The standard memory test showed 21 % errors (Figure 3.2a), and the walking-ones test showed 28% errors (Figure 3.3a).

At 280 krads, with a monotonicity erroı of 90%, most of the synapse-neuron curves were flat at 8 volts and only a very few curves still had a sigmoidal shape. There were 86% memory errors for the standard test and 76% memory errors for the walking-ones test. All the synaptic memory bits had equal number of errors within the first 200" errors reported.

The monotonicity and memory tests both showed column dependency of the errors. The monotonicity errors showed only a slight dependency as the number of errors reflected the column number that the synapse was located at. Thus, the last 9 columns, where there were 31 synapses, had more errors than the other columns. The memory errors started for the last 9 columns at a dose of 2.()() krads while the other columns had errors starting at a dose of 280 krads.

At the final cumulative dose of 440 krads, the chip was fully nonfunctional and could not be tested due to excessive power consumption. The chip was left to anneal at room temperature. However, even after 200 hours of annealing, the chip was still drawing too much power to be tested.

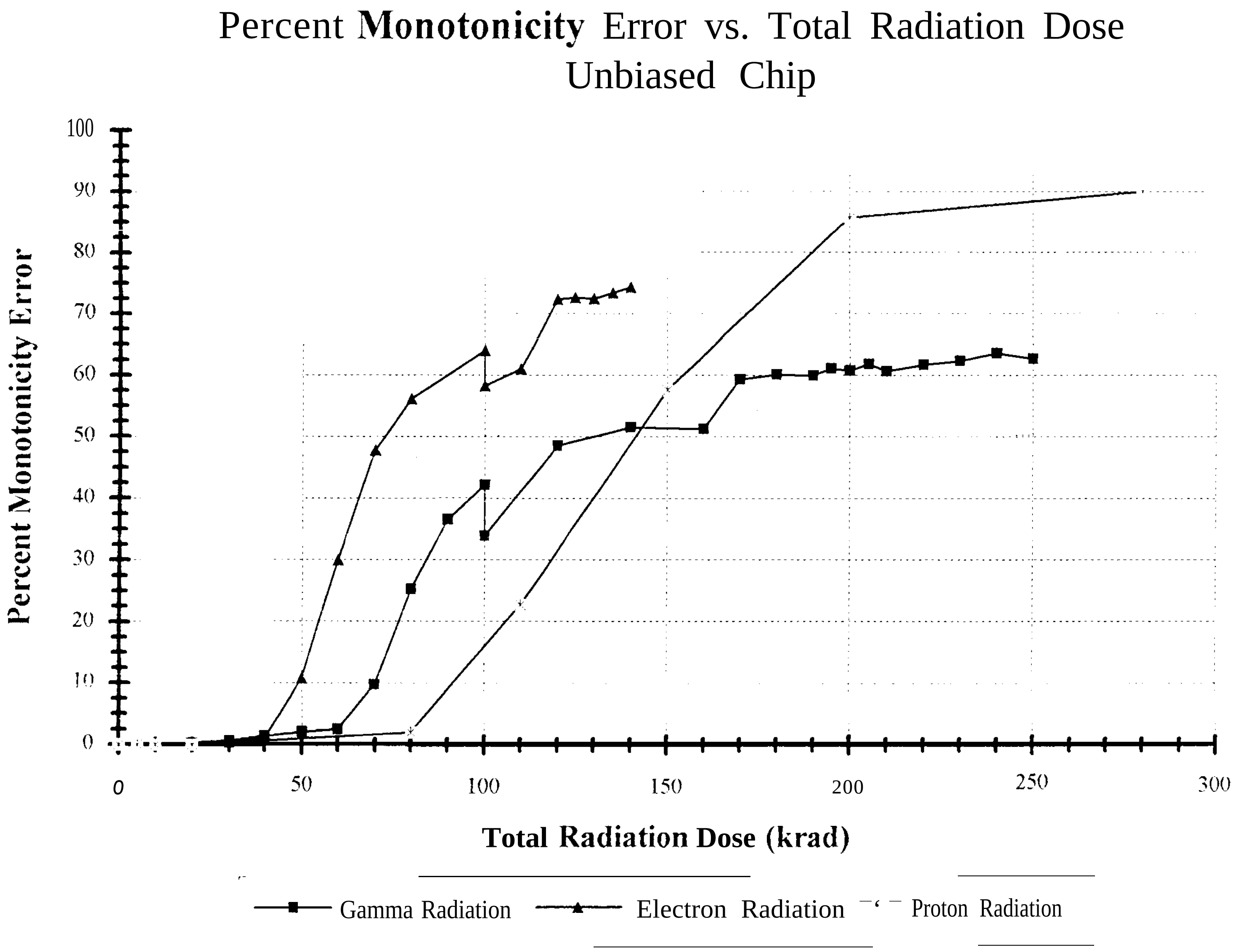

Figure 3.1a Percent Monotonicity Error vs. Total Radiation Dose for All Unbiased Chips (Gamma, Electron, and Proton Radiations)

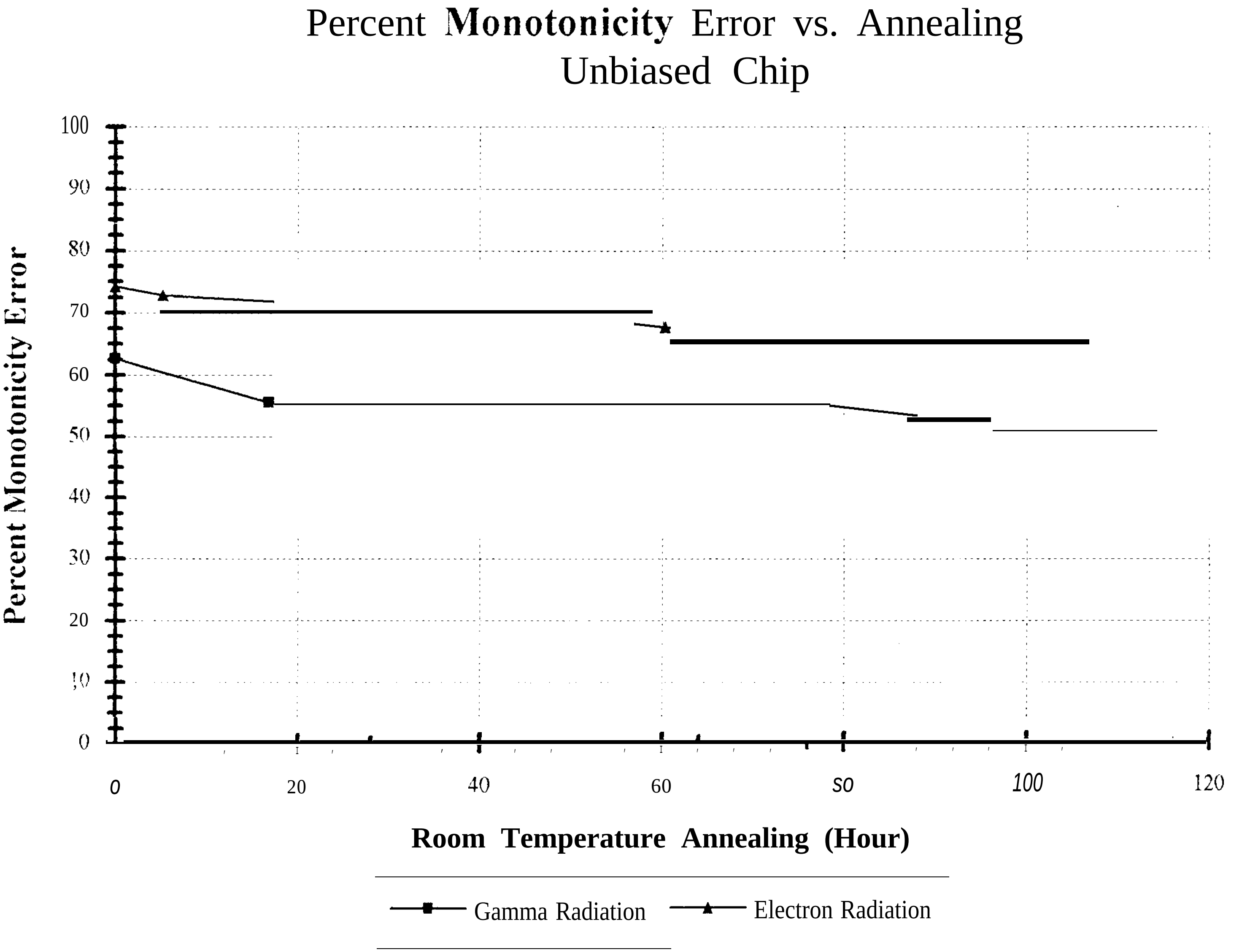

Figure 3.1b Percent Monotonicity Error vs. Annealing for All Unbiased Chips (Gamma, Electron, and Proton Radiations)

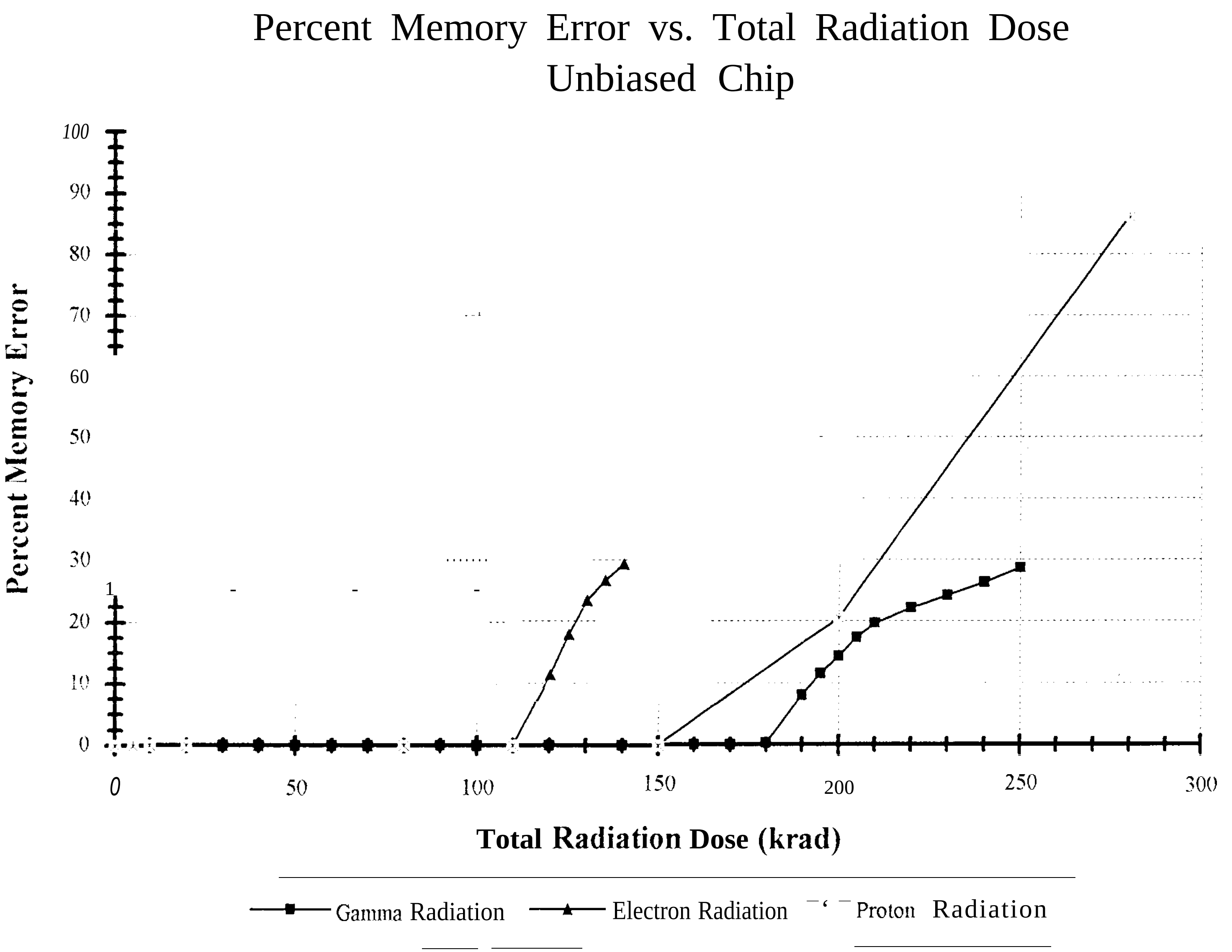

Figure 3.2a Percent Memory Error vs. Total Radiation Dose for All Unbiased Chips (Gamma, Electron, and Proton Radiations)

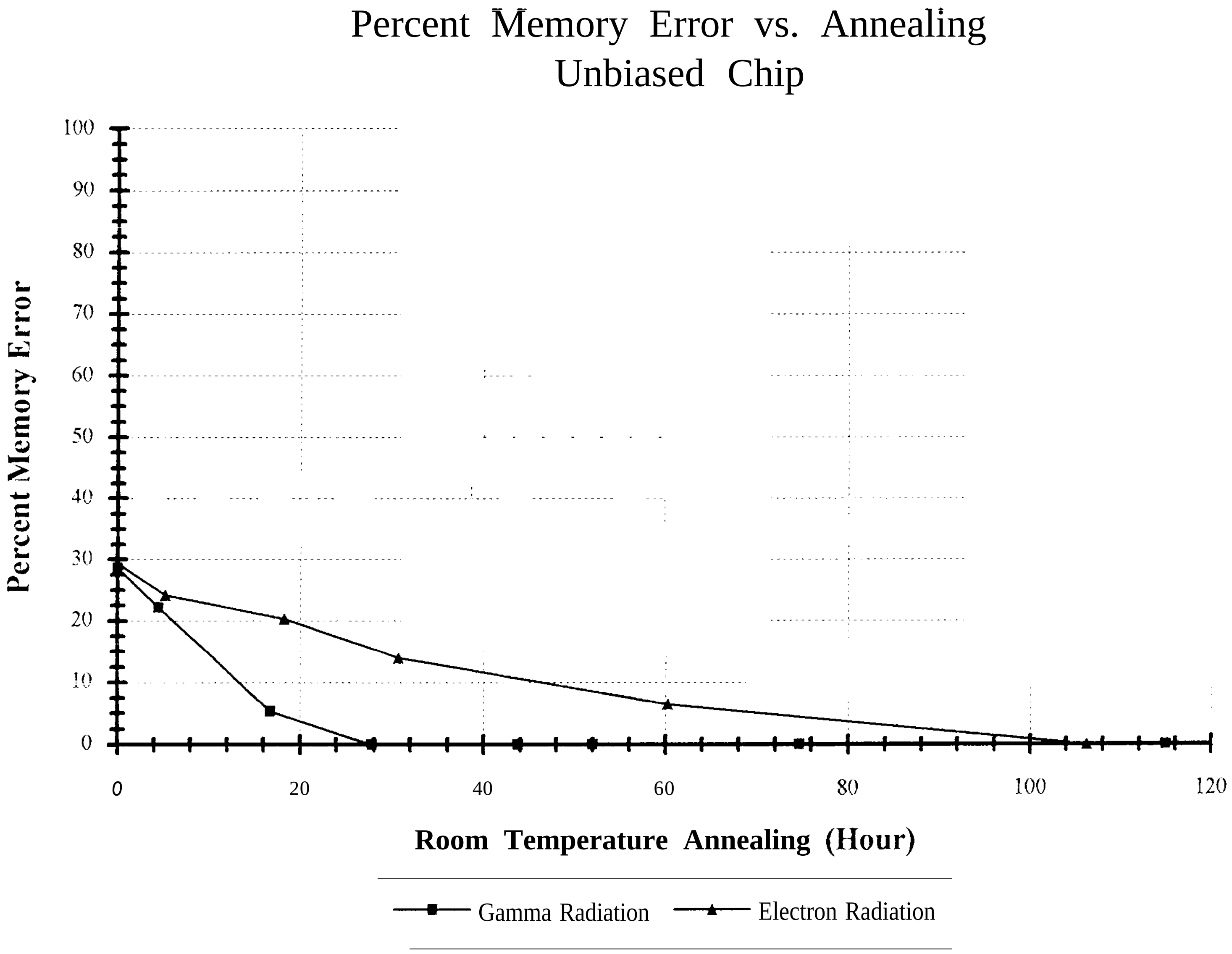

Figure 3.2b Percent Memory Error vs. Annealing for All Unbiased Chips (Gamma, Electron, and Proton Radiations)

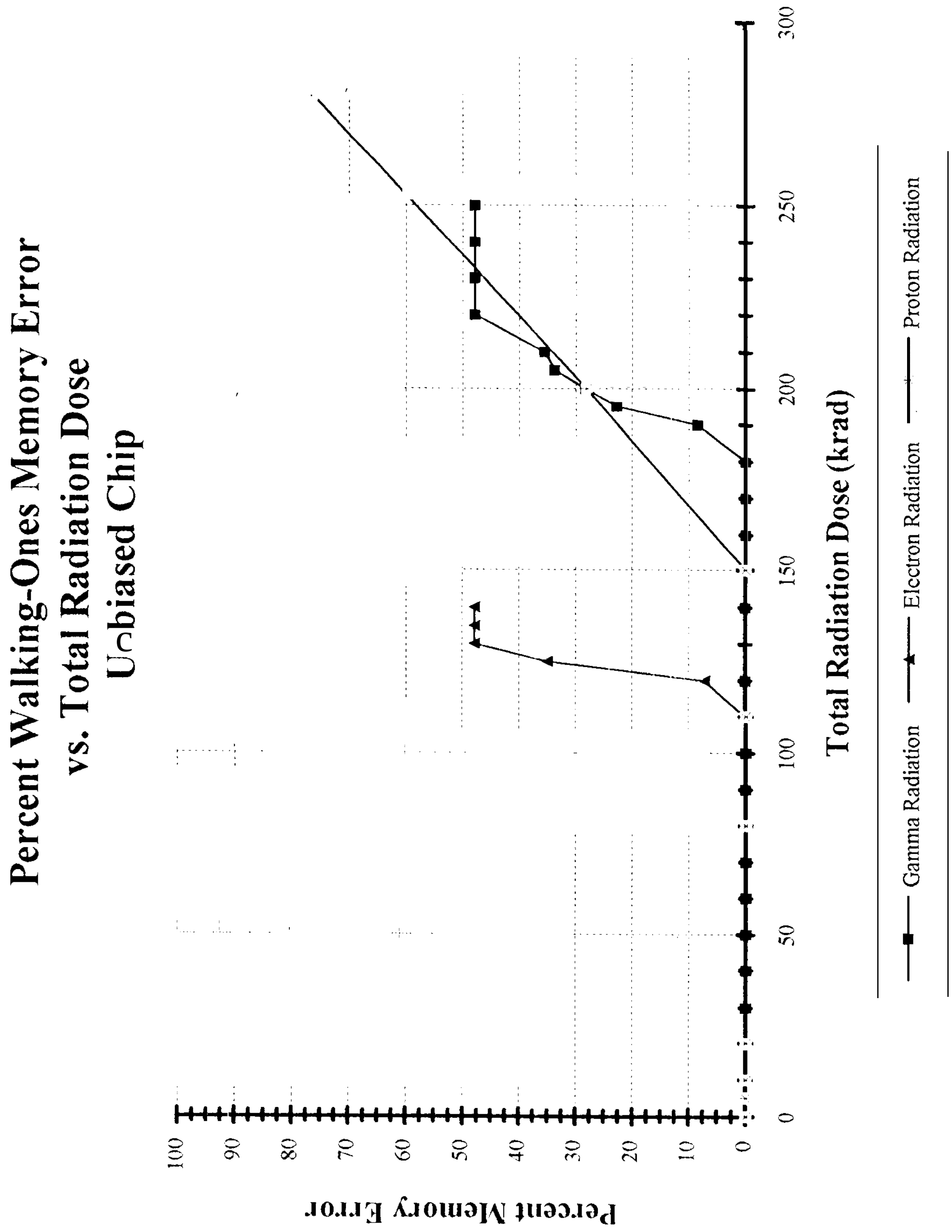

Figure 3.3a Percent Walking-Ones Memory Error vs. Total Radiation Dose for All Unbiased Chips (Gamma, Electron, and Proton Radiations)

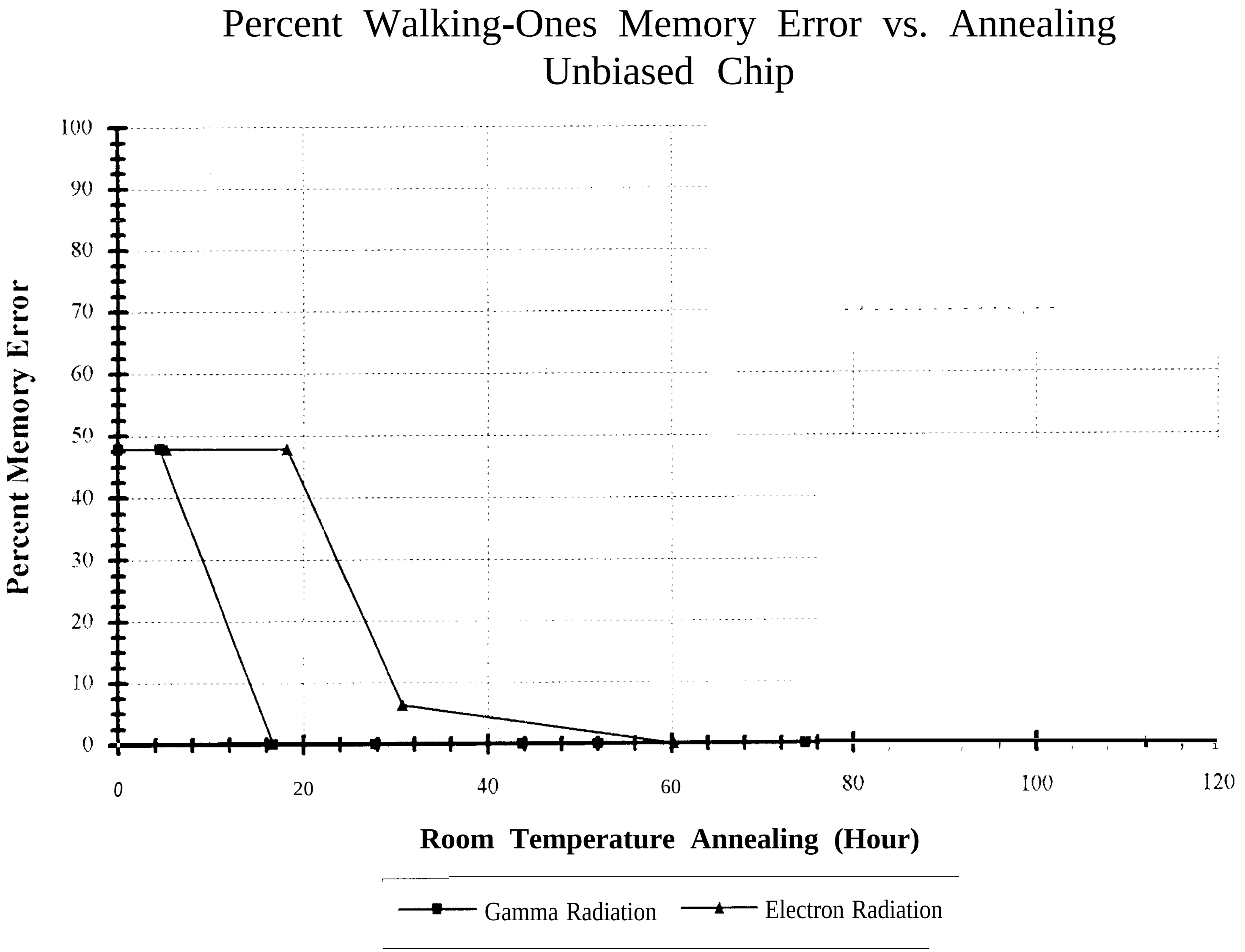

Figure 3.3b Percent Walking-Ones Memory Error vs. Annealing for All Unbiased Chips (Gamma, Electron, and Proton Radiations)

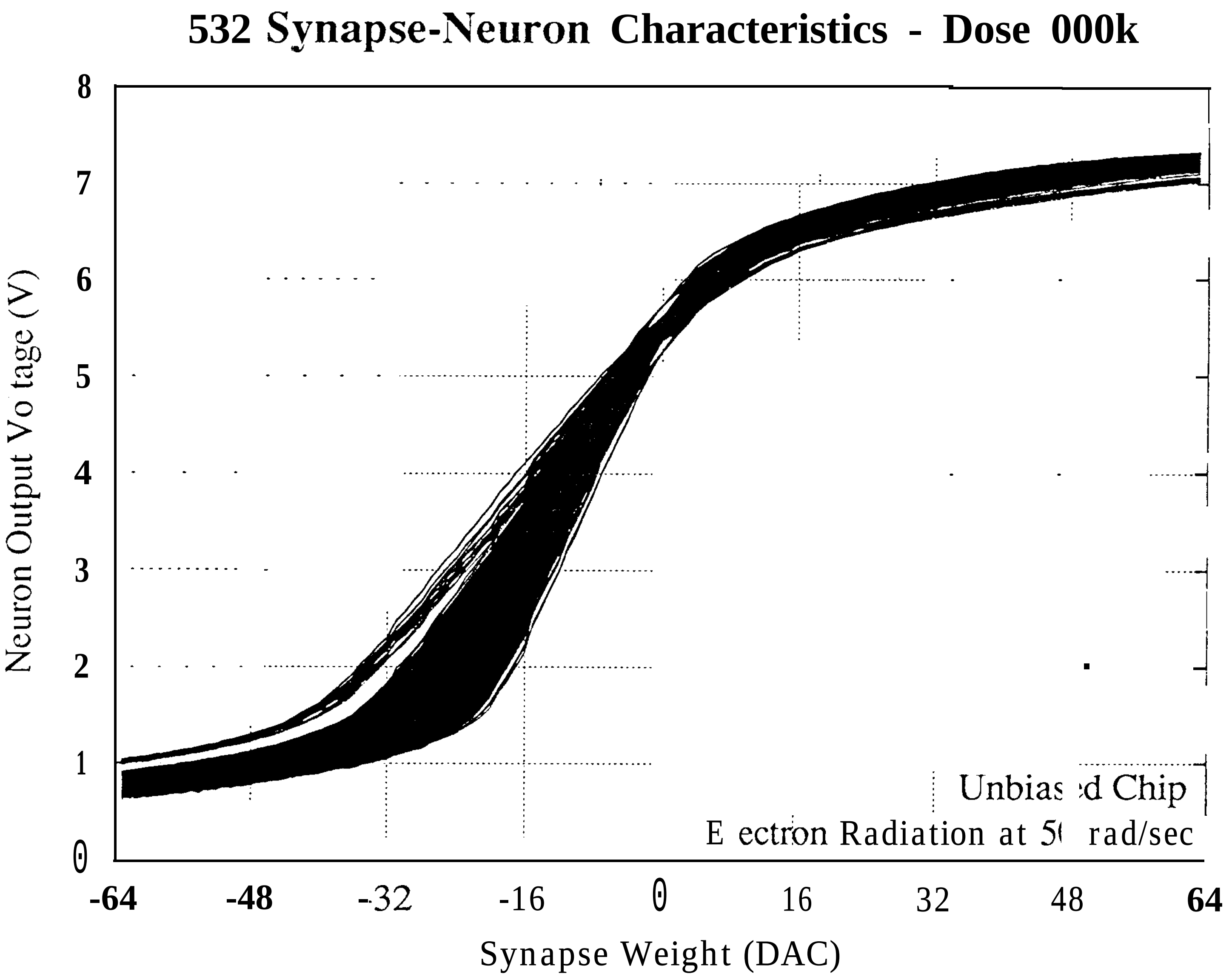

Figure 4.1a 532 Synapse-Neuron Characteristics at 0 krad Unbiased Chip under Electron Radiation

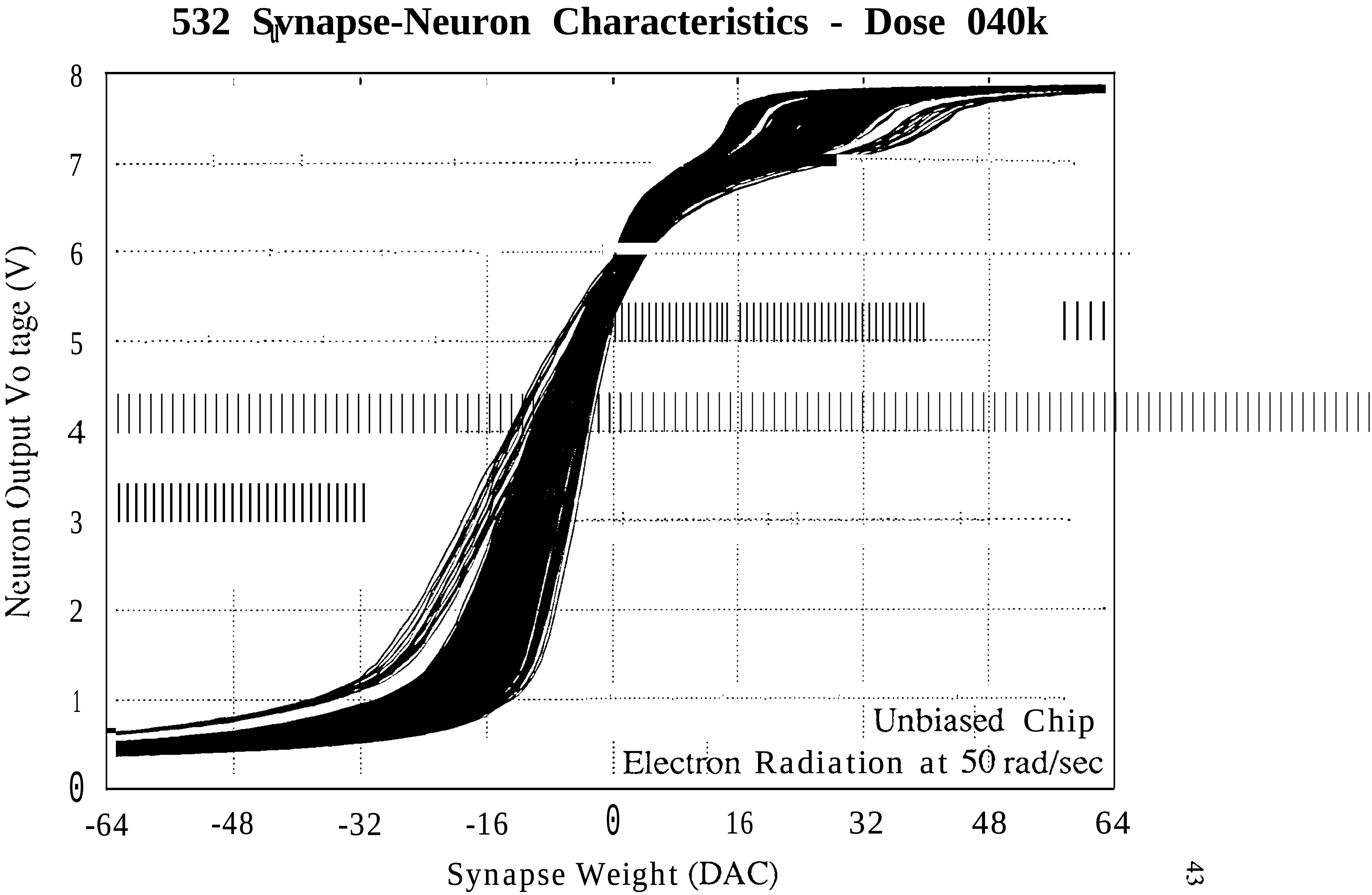

Figure 4.1b 532 Synapse-Neuron Characteristics at 40 krads Unbiased Chip under Electron Radiation

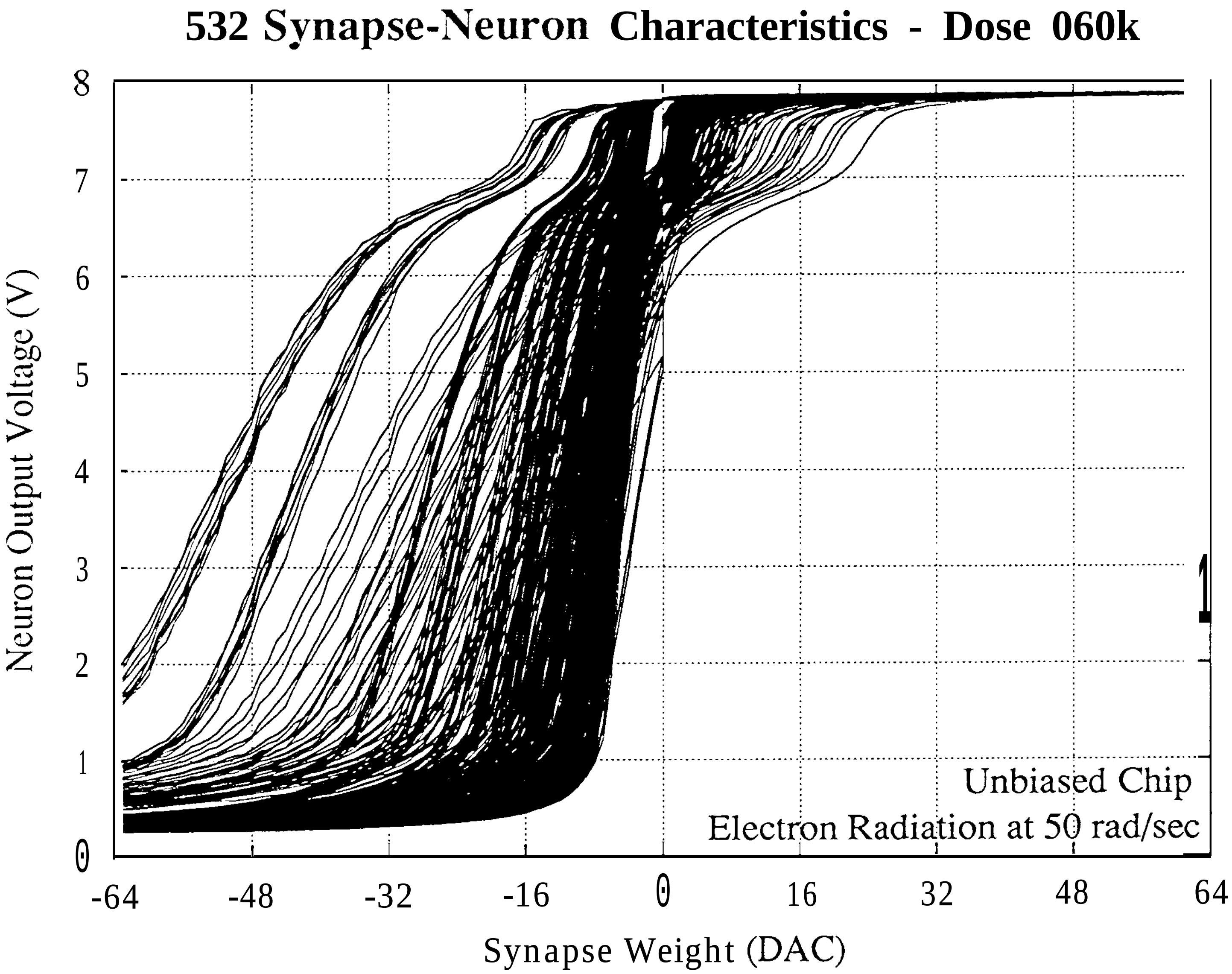

Figure 4.1c 532 Synapse-Neuron Characteristics at 60 krads
Unbiased Chip under Electron Radiation

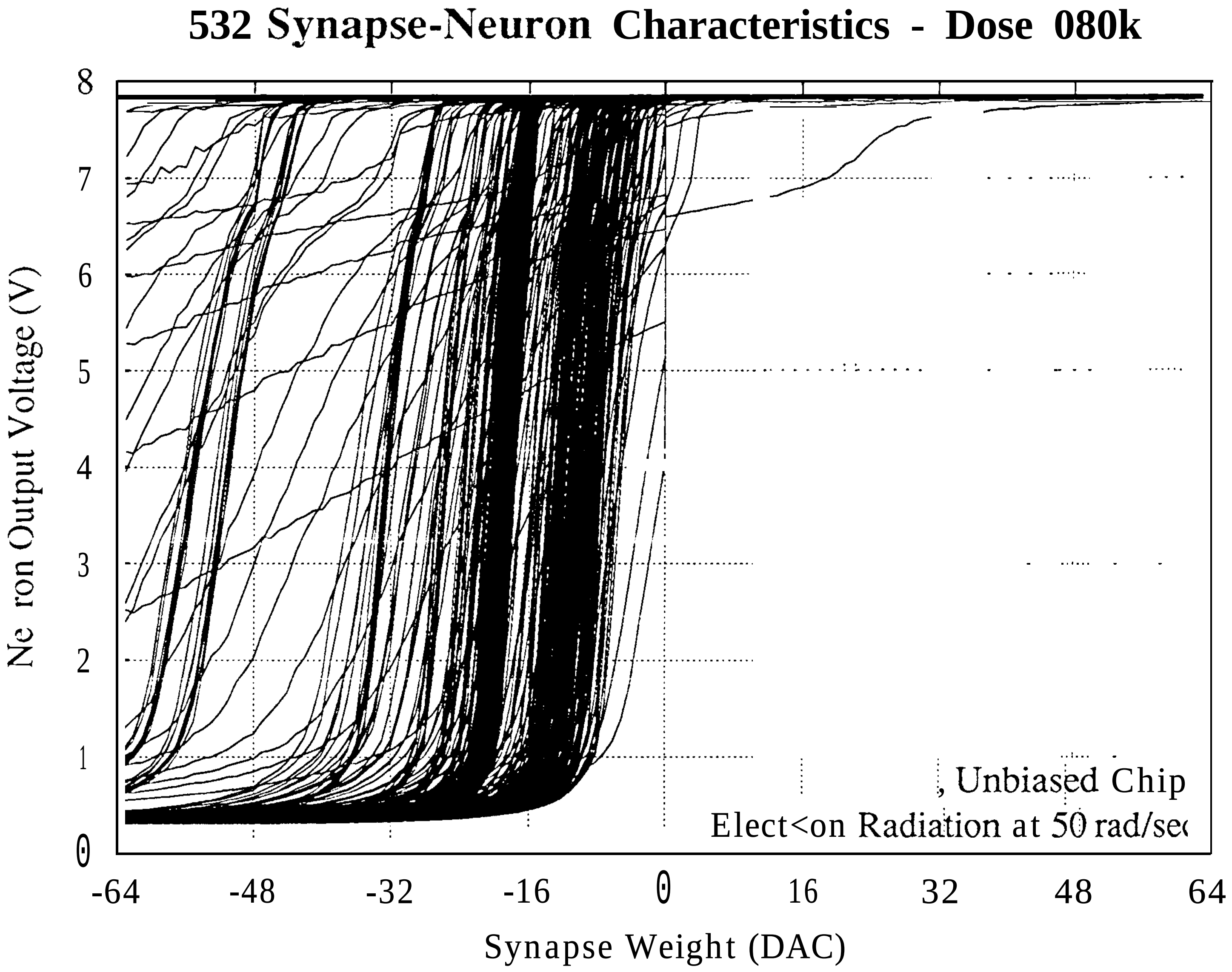

Figure 4.1d 532 Synapse-Neuron Characteristics at 80 krads Unbiased Chip under Electron Radiation

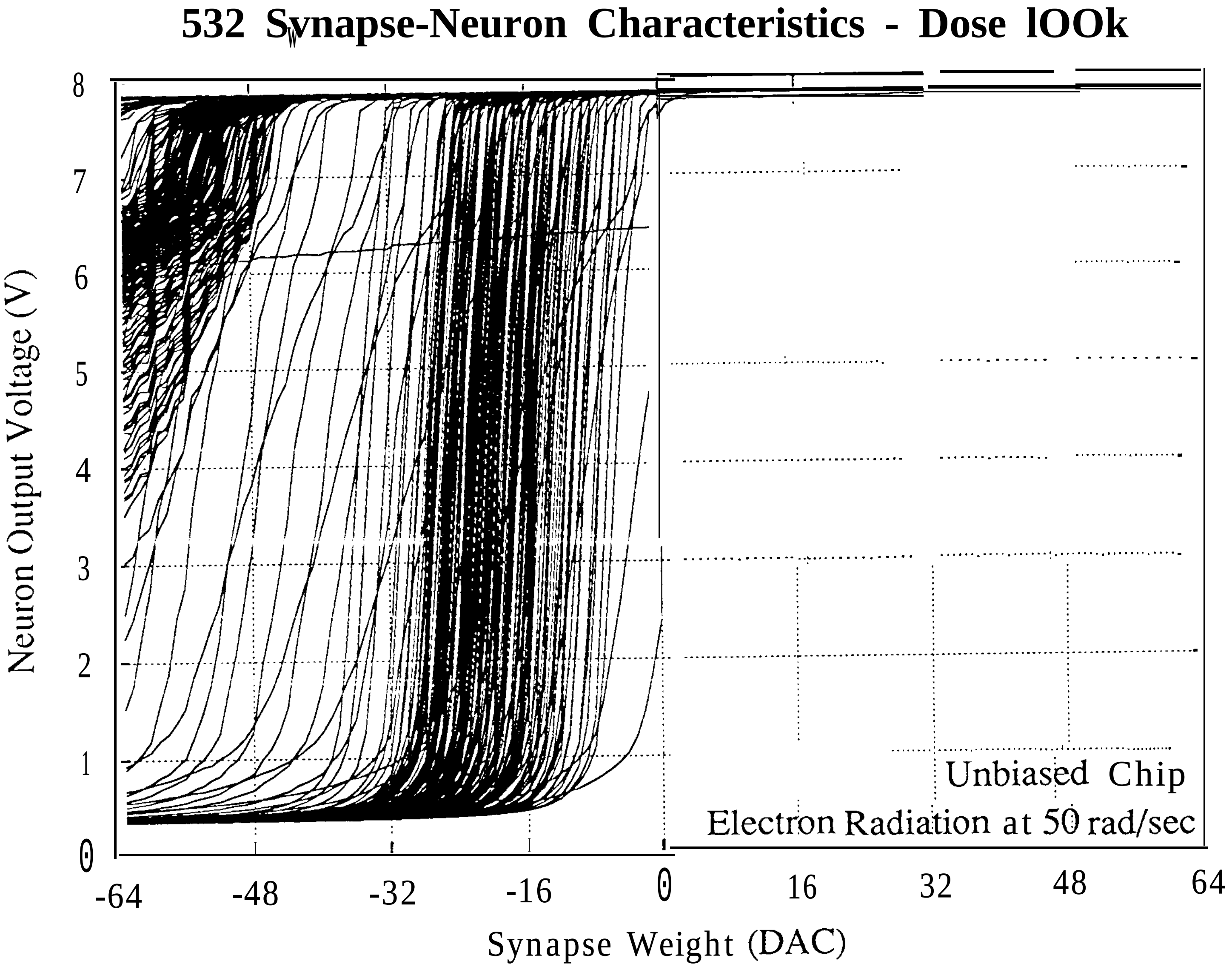

Figure 4.1e 532 Synapse-Neuron Characteristics at 100 krads Unbiased Chip under Electron Radiation

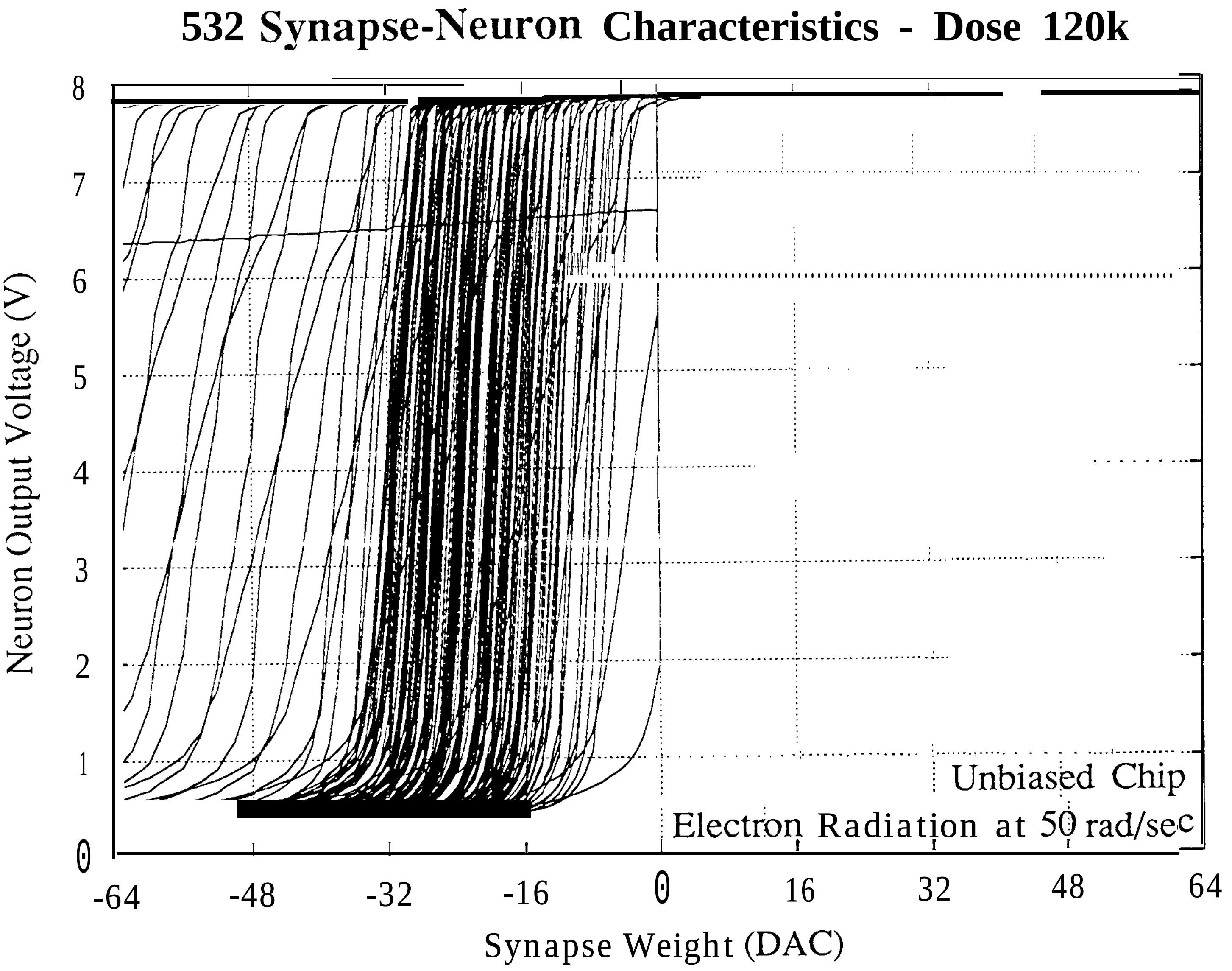

Figure 4.1f 532 Synapse-Neuron Characteristics at 120 krads Unbiased Chip under Electron Radiation

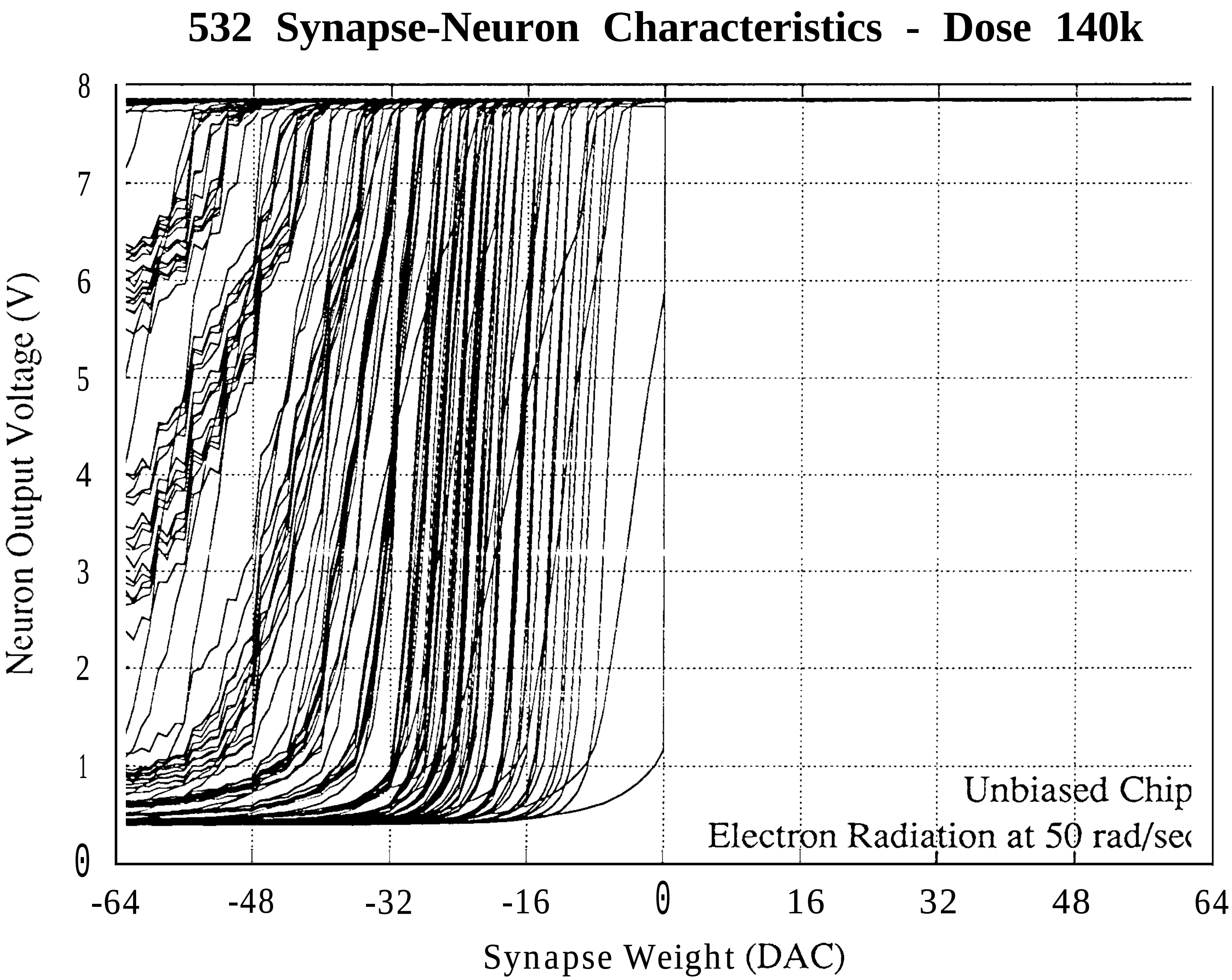

Figure 4.1g 532 Synapse-Neuron Characteristics at 140 krads Unbiased Chip under Electron Radiation

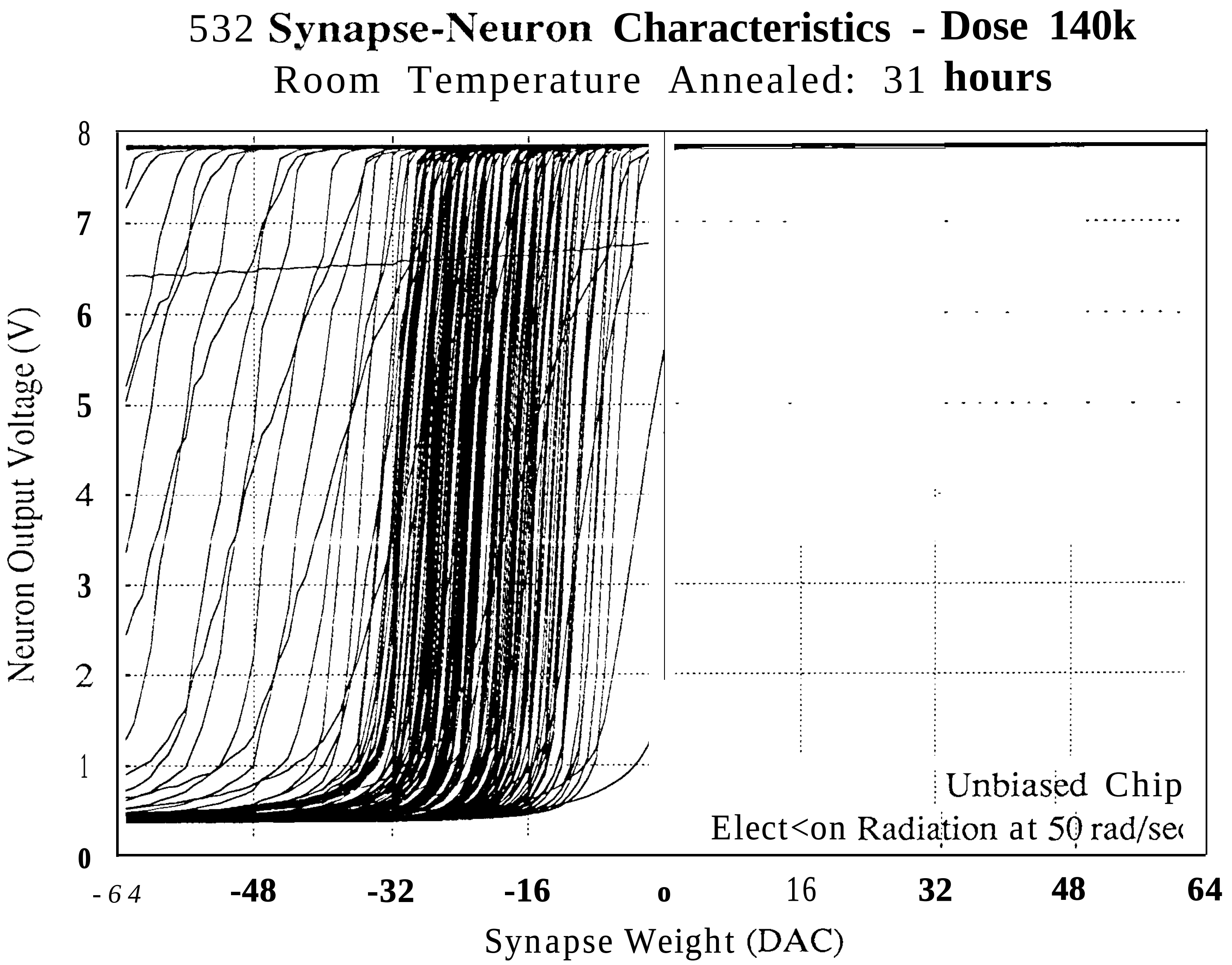

Figure 4.1h 532 Synapse-Neuron Characteristics at 140 krads, Annealed 31 hours Unbiased Chip under Electron Radiation

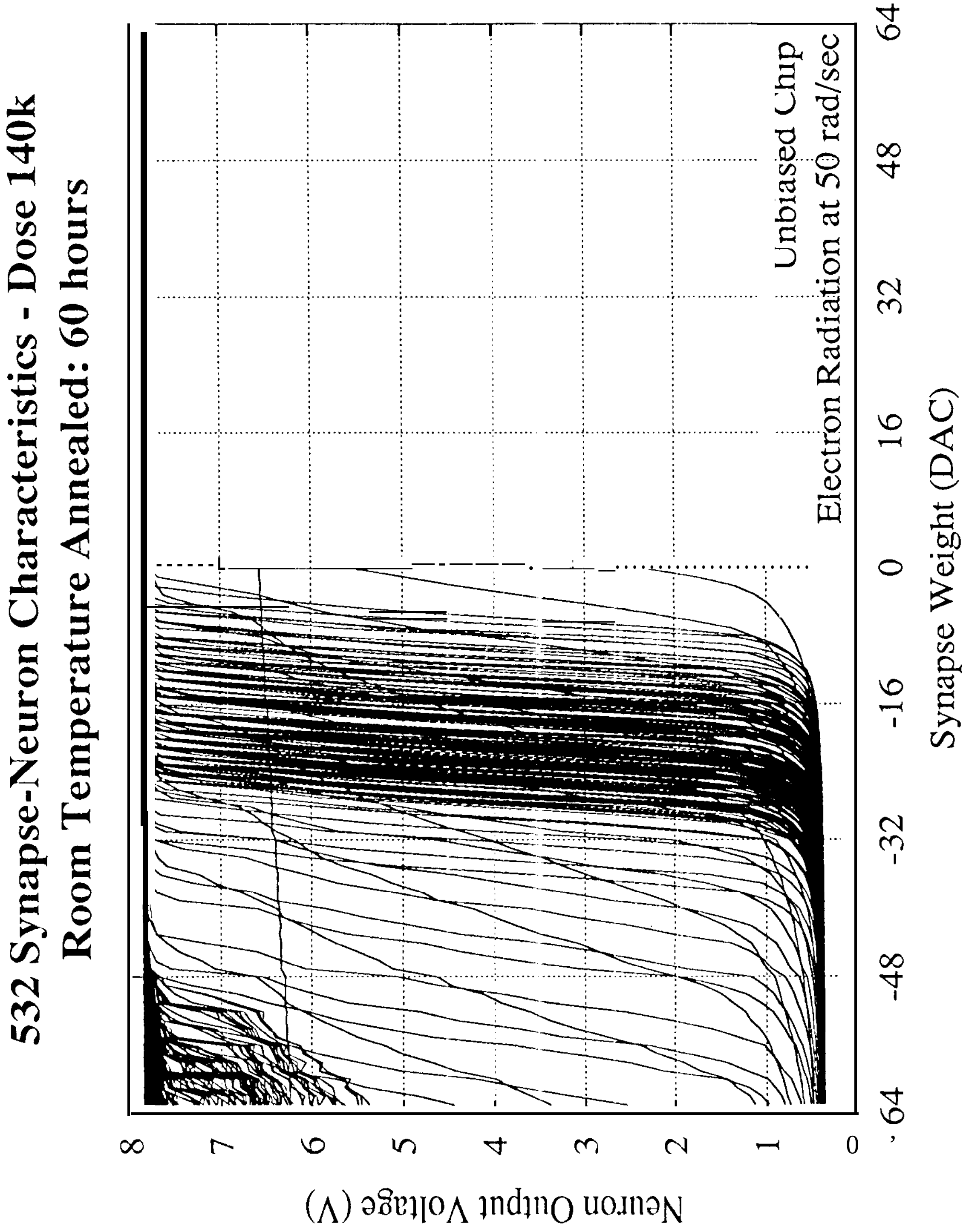

Figure 4.1i 532 Synapse-Neuron Characteristics at 140 krads, Annealed 60 hours Unbiased Chip under Electron Radiation

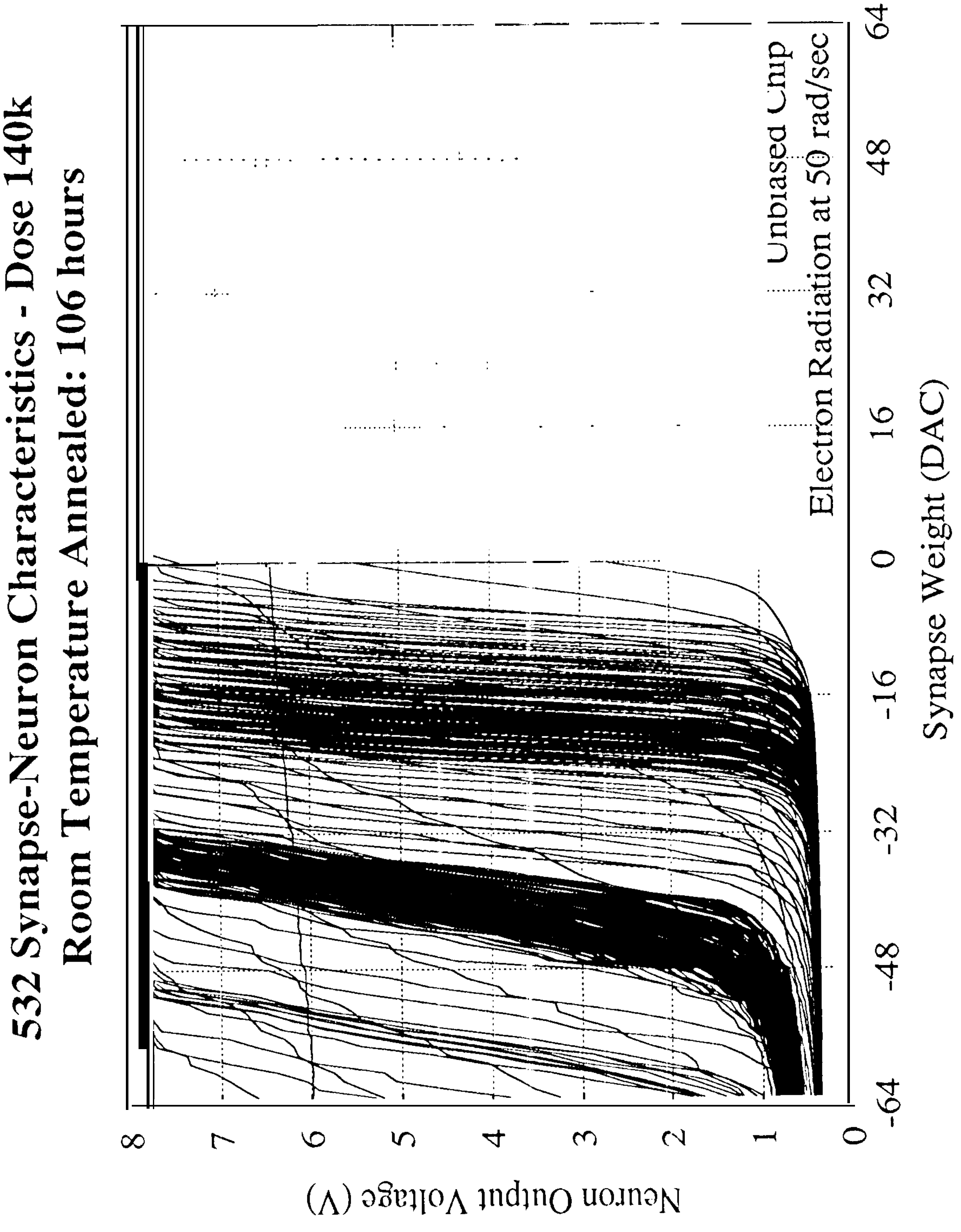

Figure 4.1j 532 Synapse-Neuron Characteristics at 140 krads, Annealed 106 hours Unbiased Chip under Electron Radiation

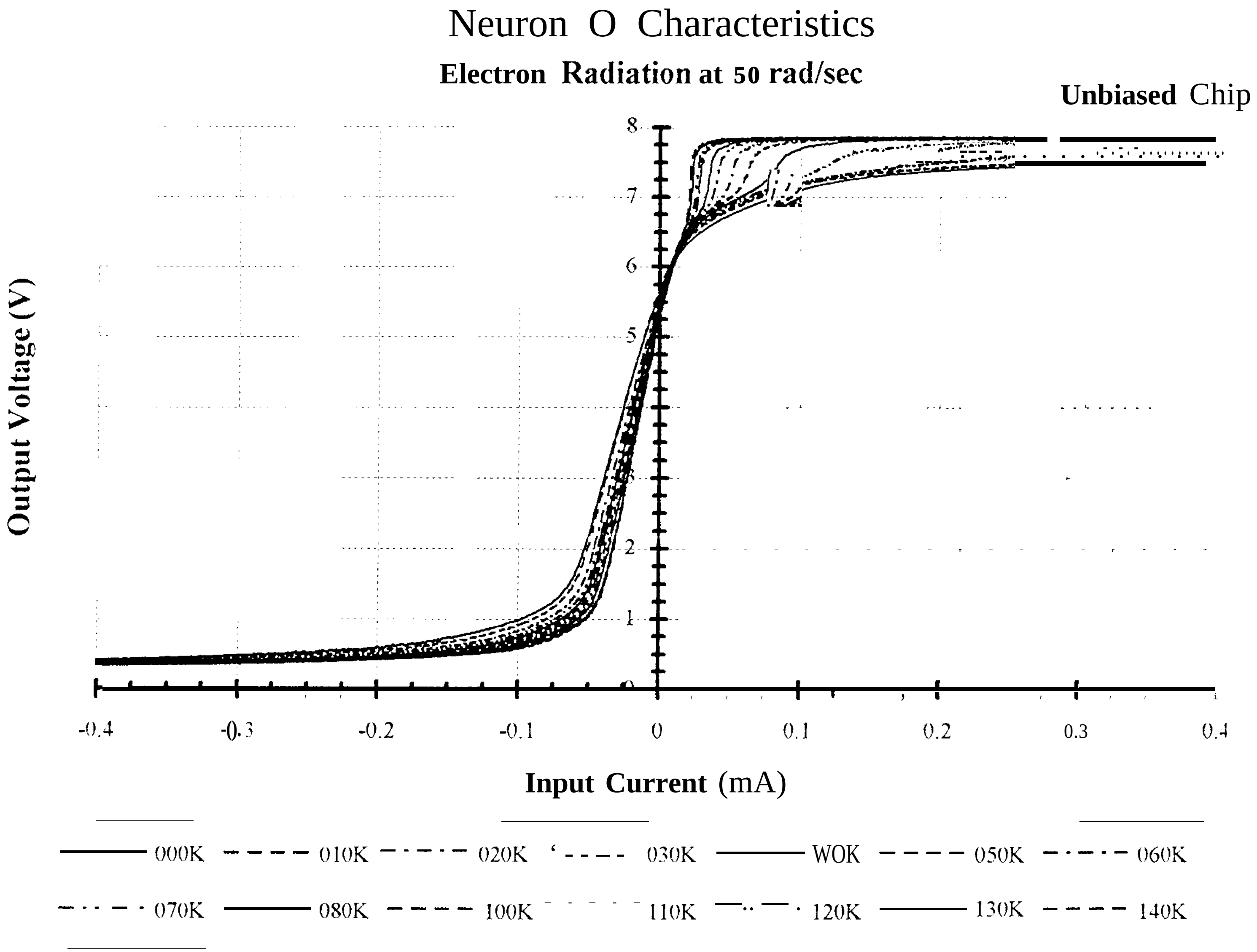

Figure 4.2 Neuron 0 Characteristics
Unbiased Chip under Electron Radiation

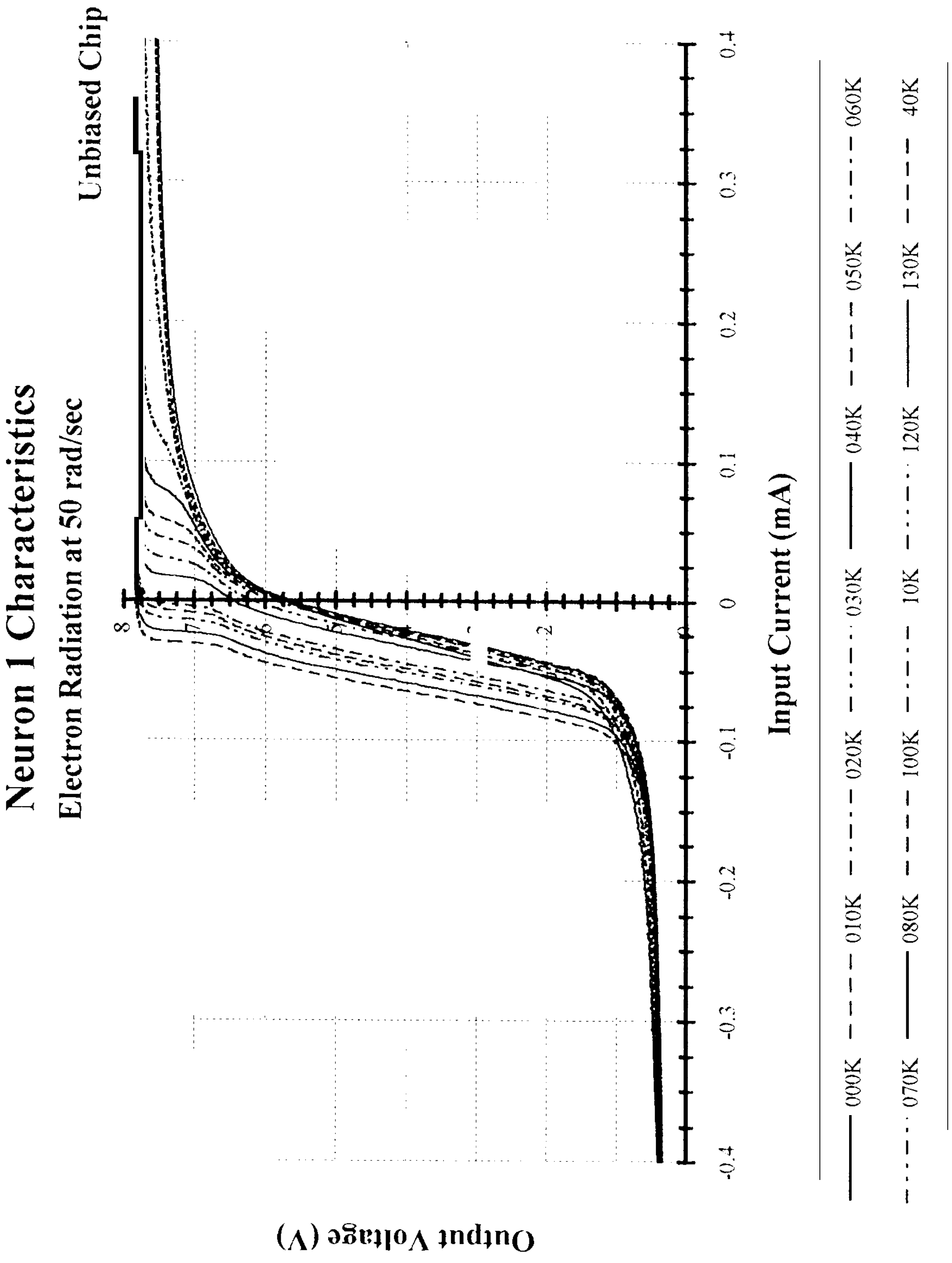

Figure 4.3 Neuron 1 Characteristics
Unbiased Chip under Electron Radiation

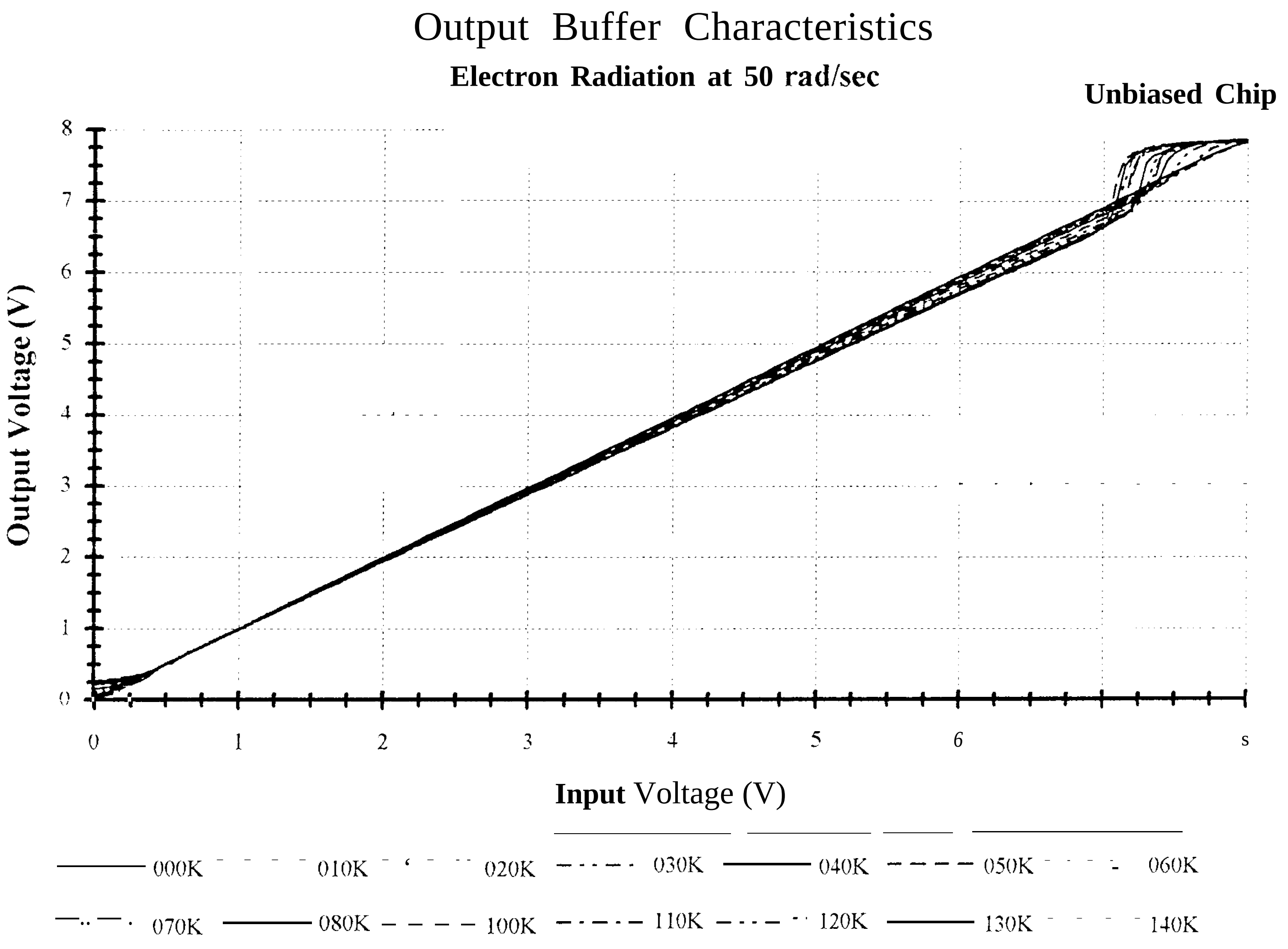

*Figure 4.4 Output Buffer Characteristics
Unbiased Chip under Electron Radiation*

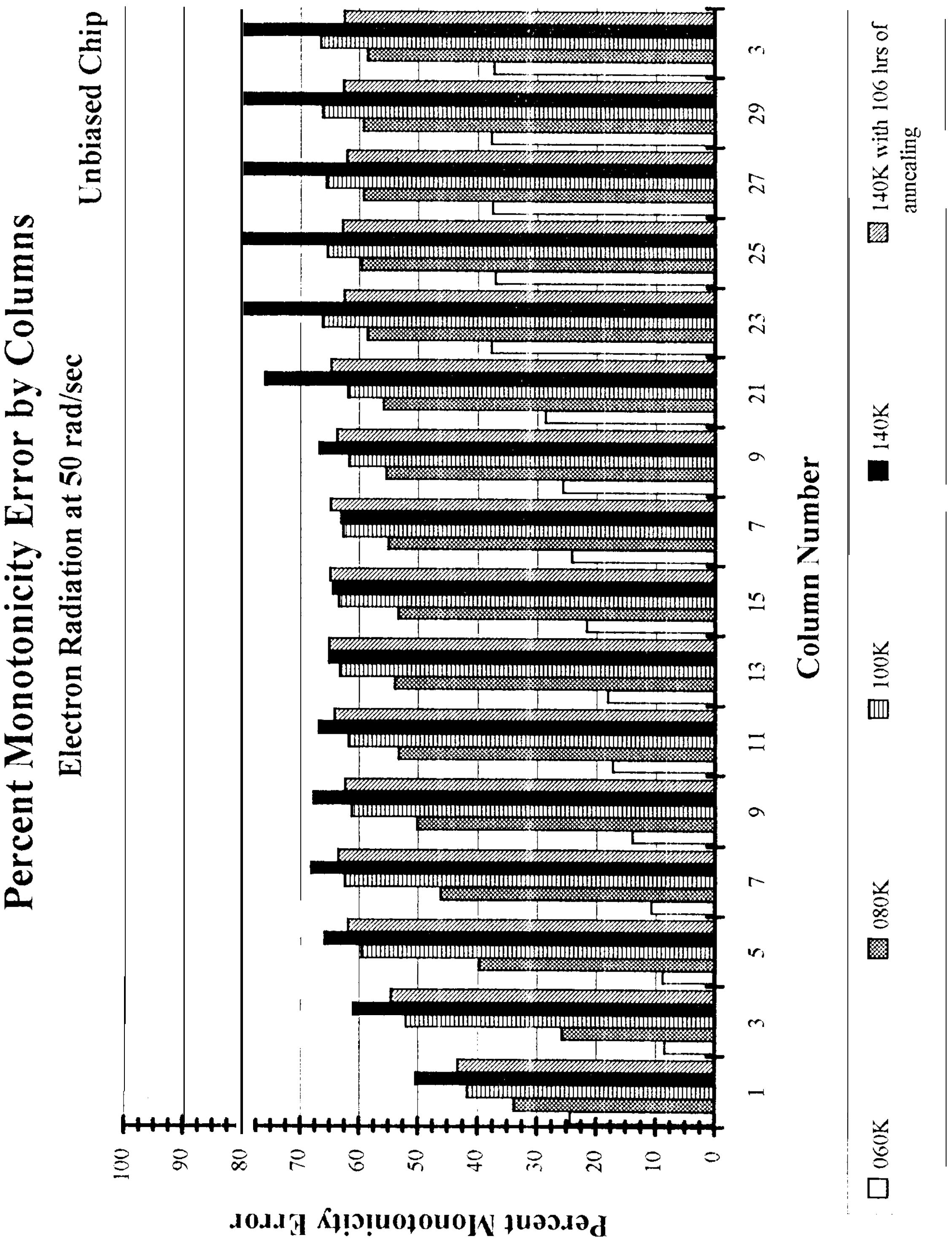

Figure 4.5 Percent Monotonicity Error by Columns Unbiased Chip under Electron Radiation

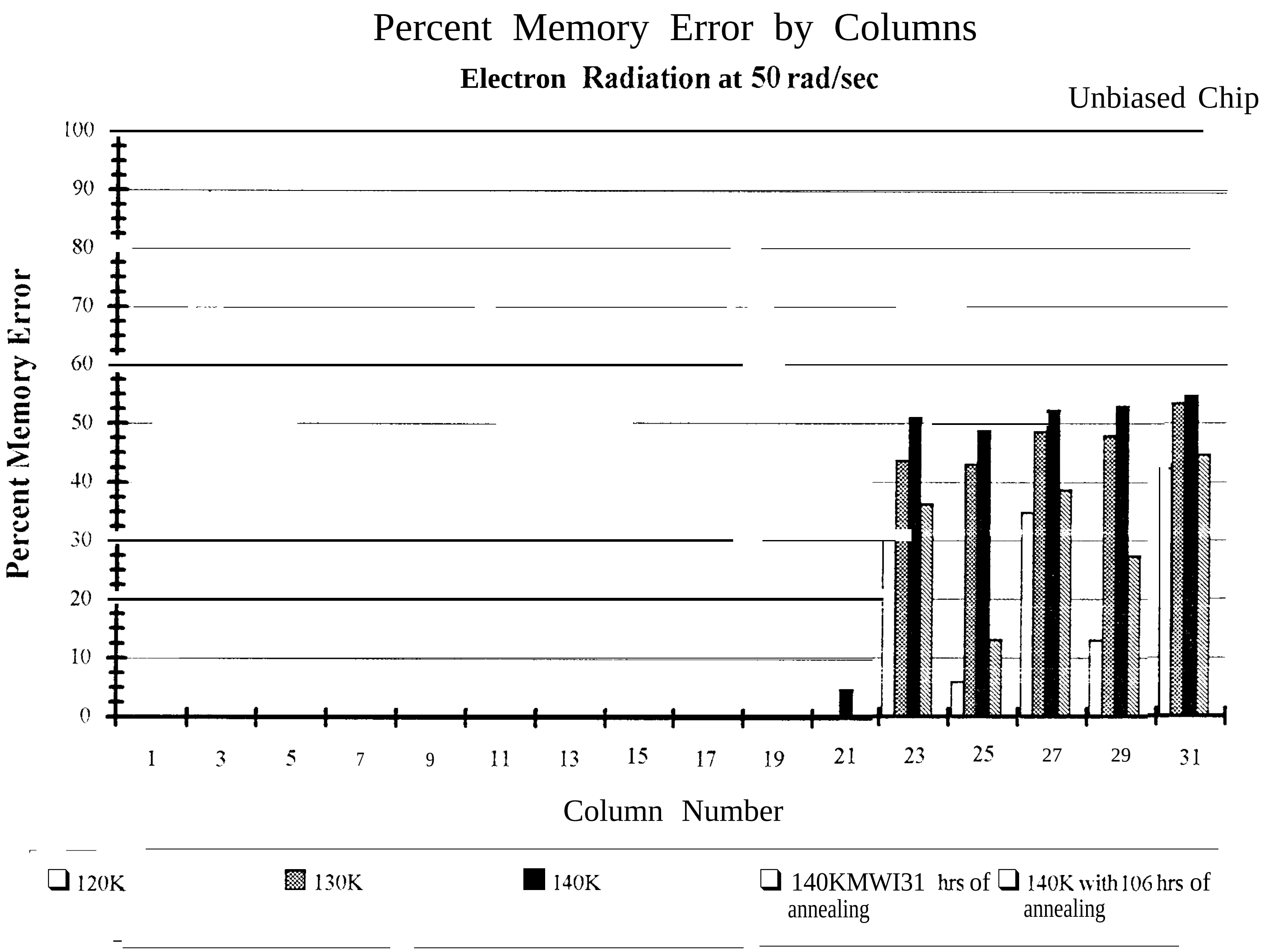

Figure 4.6 Percent Memory Errors by Columns Unbiased Chip under Electron Radiation

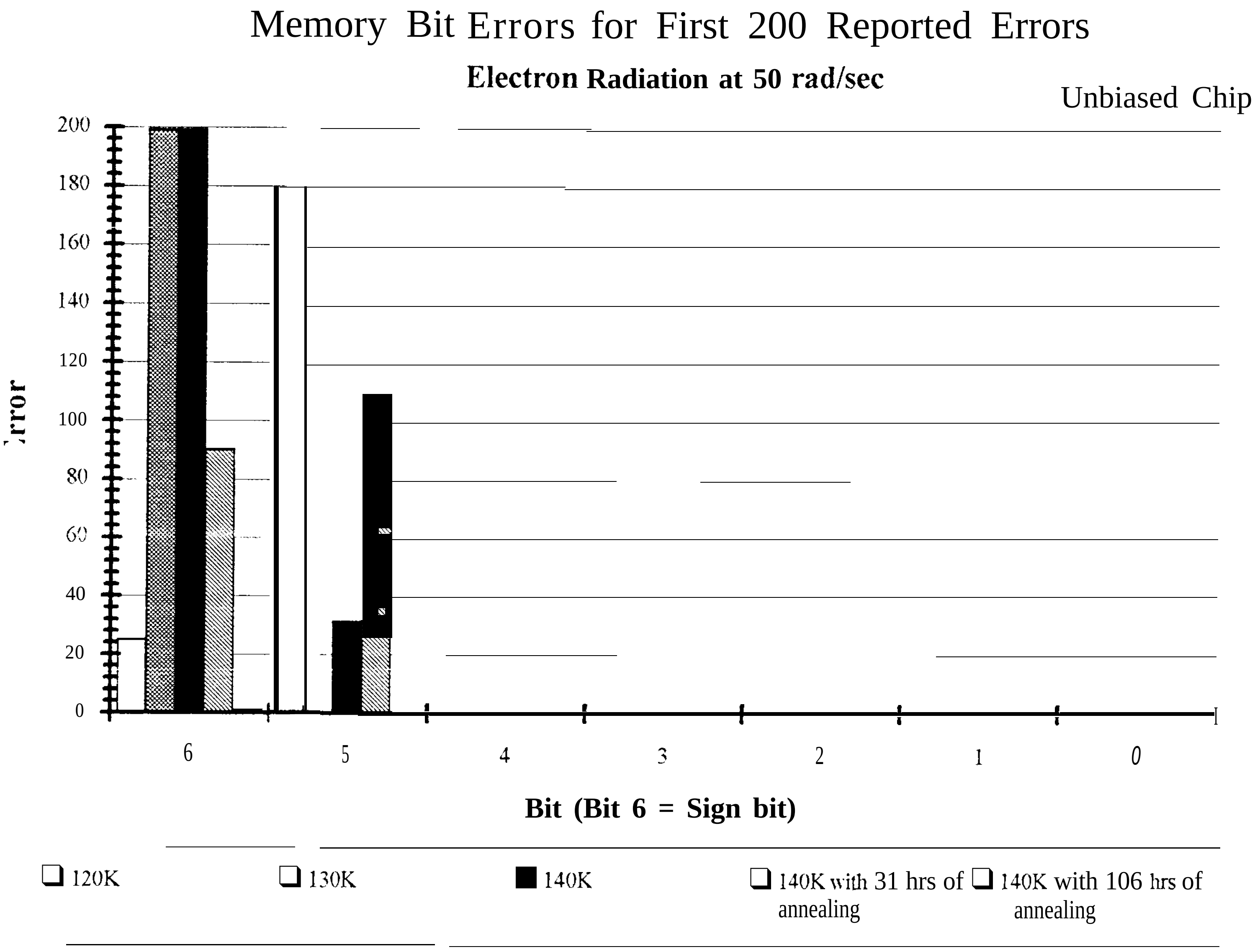

Figure 4.7 Memory Bit Errors for First 200 Reported Errors Unbiased Chip under Electron Radiation

4.2 Biased Chips

All the biased chips, exposed to electron, gamma (low or high dose rate), or *proton* radiations, showed no change in the shape of the sigmoidal curves for the neuron () and 1 tests for any of the cumulative doses. This is shown in Figure 6.2 and 6.3 for the electron radiation and in Figure *7.2* and 7.3 for the low dose rate gamma radiation. This indicates that there was no degradation or radiation effect to the neuron circuit itself. The neuron 1 test, however, showed the sigmoidal curve shifting slightly to the left as the cumulative radiation dose increased. Since neuron 1 also has input current coming from one synapse, the left shift maybe attributed to the leakage current coming from that synapse as it degraded with radiation.

The output buffer test also showed no change in the linearity between the input and output voltages as the cumulative radiation dose was increased for all the biased chips. Figure 6.4 and 7.4 show the output buffer characteristics from the electron and gamma (low dose rate) radiation experiments. '1'bus, the main degradation for the biased chips was from the synapses where the synapse-neuron curves showed a left shift.

Room temperature annealing was done following exposure for all the biased chips. The synapse-neuron curves did not show any annealing even after 100 hours, and the monotonicity errors remained unchanged, The percent monotonicity errors during annealing is shown in Figure 5.1b for all the biased chips under the three different radiation sources. '1'here was some variation to the monotonicity errors as the electron radiation, followed by gamma radiation at low and high dose rates and proton radiation. Figures are only provided for the electron and low dose rate gamma radiations. The other radiation sources showed similar effects as can be compared in Figures 5.1a, 5.2a, and 5.3a where the percent monotonicity errors, percent memory errors, and percent walki rig-ones memory errors versus total radiation dose, respectively, arc shown.

4.2.1 Electron Radiation

The 532 synapse-neuron characteristics of the chip at () krads is shown in Figure 6.1 a The chip performed without any effects when tested at 2 krads. The symipse-neuron curves became steeper and shifted to the left as exhibited in Figure 6.1 b, causing 16% monotonicity errors (Figure 5.1a) at 4 krads. With 70% monotonicity errors at 5 krads, most of the synapse-neuron curves (Figure 6.1 c) were fiat at 8 volts (neurons output high) for all the input synapse weight values. Only a few synapse-neuron curves still showed a sigmoidal curve. For both the 4 krad and 5 krad doses, there were no memory errors for either the standard (Figure 5.2a) or the walking-ones (Figure 5.3a) tests.

Finally, at 6 krads, the chip latched-up and drew excessive power during the synapse-neuron and monotonicity tests. Memor y tests were still able to be performed and showed that there were 26% errors for the standard test and 22% for the walking-ones test. Figure 6.7 reveals that the synapse memory bit 0 showed significant errors in the first 200 errors reported.

Both the monotonicit y and memory tests had column dependency as portrayed in Figures *6.5 and 6.6,* respectively. For example, at 4 krads, monotonicity errors ranged from about *40%* in the first few columns to 78% in the last few columns. Memory errors at 6 krads occurred significantly only in the last 9 columns.

4.2.2 Gamma Radiation (Low dose rate)

For the gamma radiation with the low dose rate (().1 ra(i/see), the biased chip had cumulative doses of 6, 7, and 7.5 krads. Figure 7.1a shows the initial characteristics of the synapse-neuron curves with no radiation exposure. After the first radiation dose of 6 krads, the chip already started to show some degradation in the synapse-neuron curves (Figure 7.1b) and the memory test s. *'1'here were* 5% monotonicity errors (Figure 5.1a) due to the synapse-neuron curves becoming steeper and shifting to the left. The standard memory test (Figure 5.2a) had 28% errors, and the walking-ones test (Figure 5.3a) had 48% errors.

These radiation effects increased at 7 krads showing 15% monotonicity errors, 51 % errors for the standard memory test, and 71 % errors for the walking-ones memory test, The synapse-neuron curves are shown in Figure 7.1c. At the final dose of 7.5 krads, more synapse-neuron curves had shifted to the left [Figure 7.1d) with 23% monotonici ty errors. The memo ry errors increased as well to 60% for the standard test and *78%* for the walking-ones test. The sign bit (bit 6) for the synapse memory showed significant errors in both the tests for the first 200 reported errors. This can be seen in Figure 7.7. Memory bit0 also had some errors.

Co] umn dependent errors were observed for both the monotonicit y and memory tests ill Figures *7.5* and *7.6,* respectively. The column dependency was less pronounced in the monotonicity test where the errors ranged from 10% in the first few columns to 27(% in the last few columns. There were more memory errors in the last 9 columns showing 83% errors for the standard test and 90% errors for the walking-ones test compared to 30% errors for the standard test and 47% errors for the walking-ones test in the other columns at *7.5* krads. The first 5 columns did not have any memory errors .

4.2.3 Gamma Radiation (High dose rate)

For the high dose rate *(50 tad/see)* gamma radiation, the biased chip had cumulative doses of 4, 5, and 6 krads. At 4 krads, the chip did not show any signs of degradation. Synapse-neuron curves started shifting left and becoming steeper at 5 krads with 12% monotonicity errors, There were 4% errors for the standard memory test and 23% errors for the walking-ones lest,

At the final dose of 6 krads, the chip was nonfunctional due to excessive power consumption and no tests could be run. After 7 hours of room temperature annealing, the power consumption of the chip decreased. The synapse-neuron curves showed significant left shift with 31% monotonicity errors. The standard memory tests showed 47% errors, and the walking-ones memory test showed *58% errors.* Memory bits O, *5, and* 6 *(sign* bit) of the synapse showed errors for both memory tests in the first **200** erroneous results.

No memory errors occurred in the first 9 columns but the other columns showed errors increasing with their column number. ‘his column dependency was also observed in the monotonicity test where the errors ranged from about 1 ()% in the lower numbered columns to about 35% in the higher numbered columns.

4.2.4 Proton Radiation

For the proton radiation, the biased chip performed well at 1 and 3 krads, and there were no signs of degradation. The shape of synapse-neuron curves showed only a slight change by starting to become steeper at 6, 8, 10, and 15 krads. The monotonicity error was still insignificant at 15 krads (().1% errors) since the synapse-neuron curves still maintained their sigmoidal shape, without any left shift. Furthermore, no memory errors were measured for both the standard and walking-ones tests.

The cumulative dose was then doubled to 30 krads, when the chip became nonfunctional by latching-up and exceeding the current limit. None of the tests could be performed and no curves could be drawn. 1 :ven after 1()() hours of annealing time the chip continued to be nonfunctional and had excessive power consumption. Thus, the biased chip was damaged very severely with 30 krads of proton radiation.

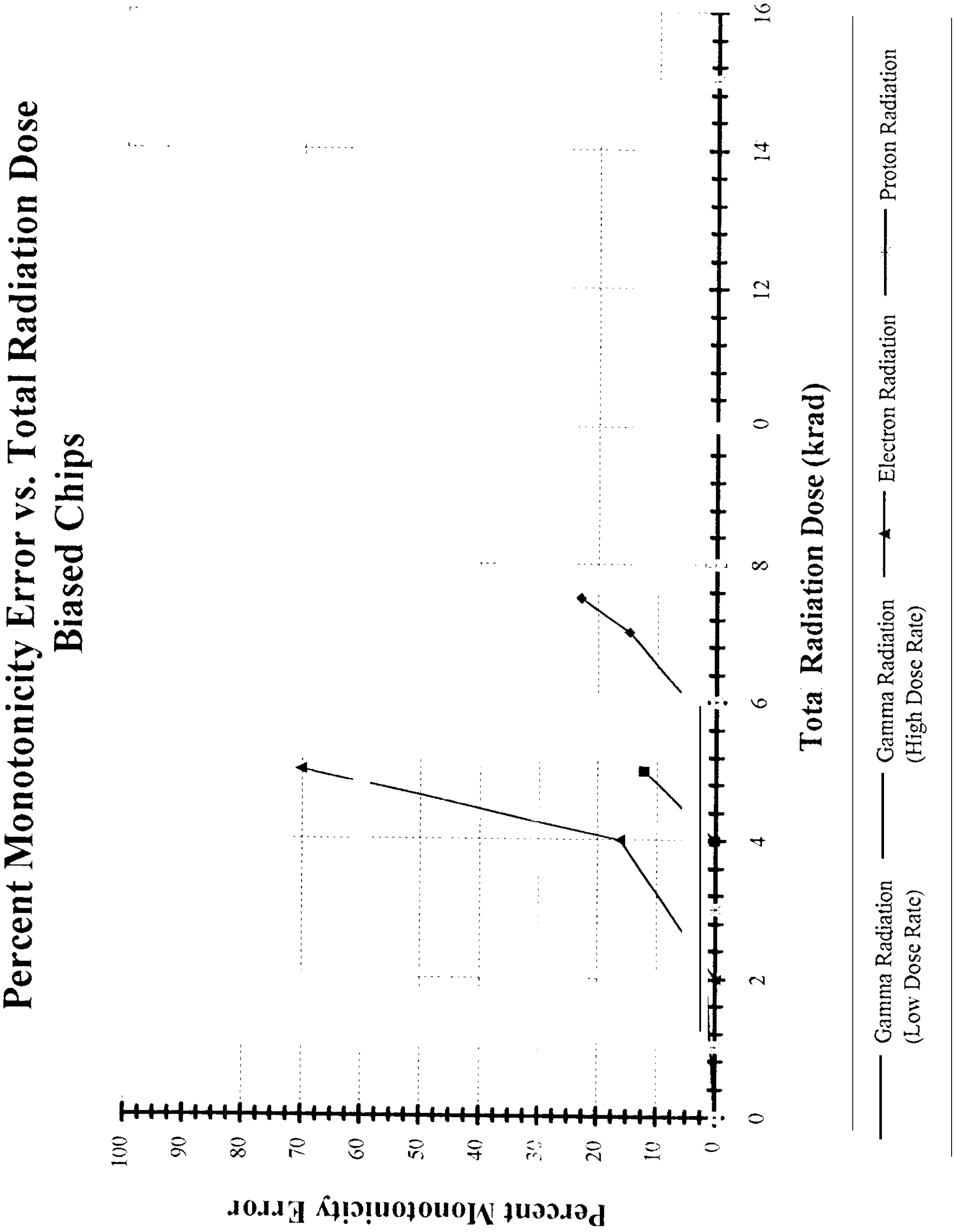

Figure 5.1a Percent Monotonicity Error vs. Total Radiation Dose for All Biased Chips (Gamma, Electron, and Proton Radiations)

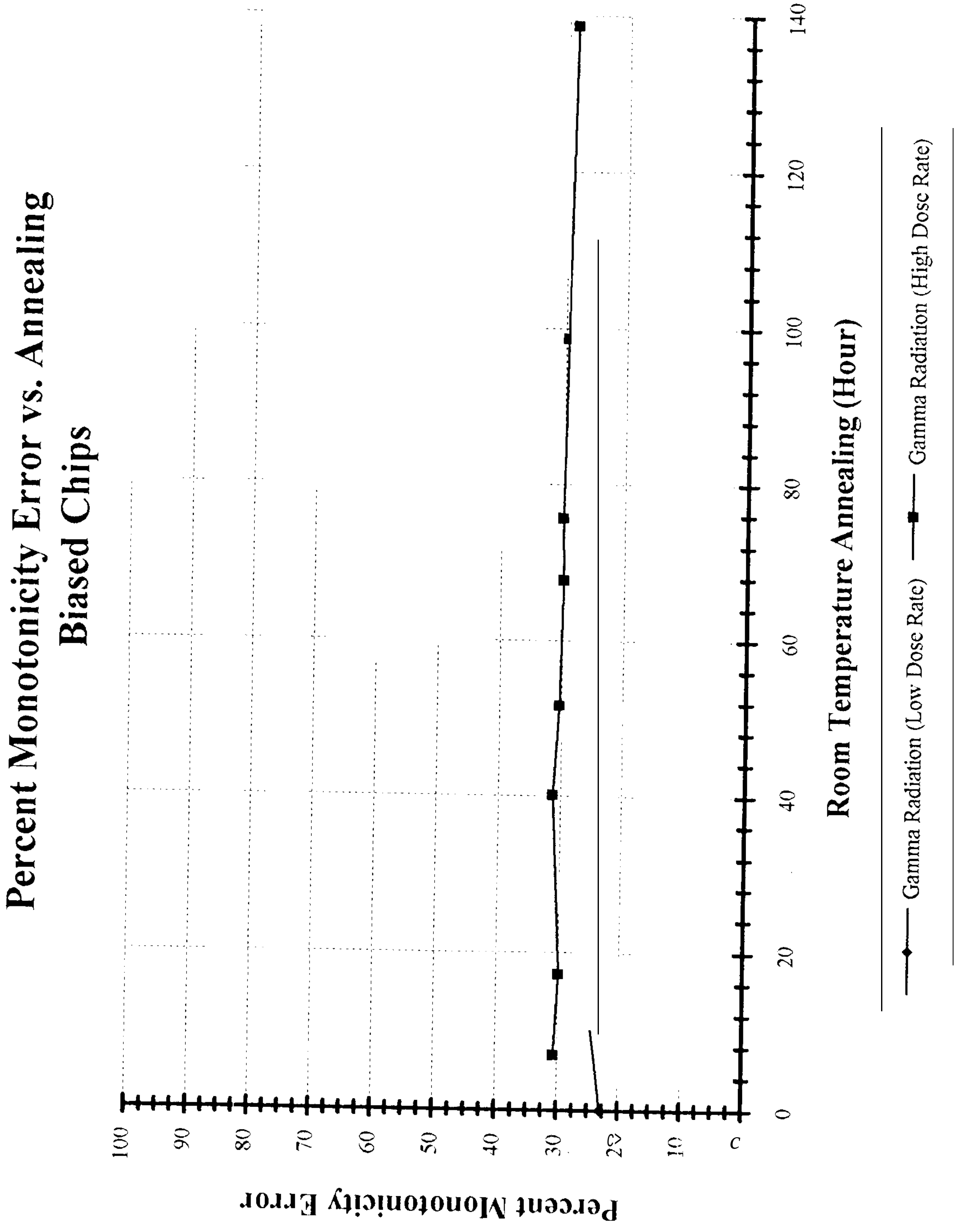

Figure 5.1b Percent Monotonicity Error vs. Annealing for All Biased Chips (Gamma, Electron, and Proton Radiations)

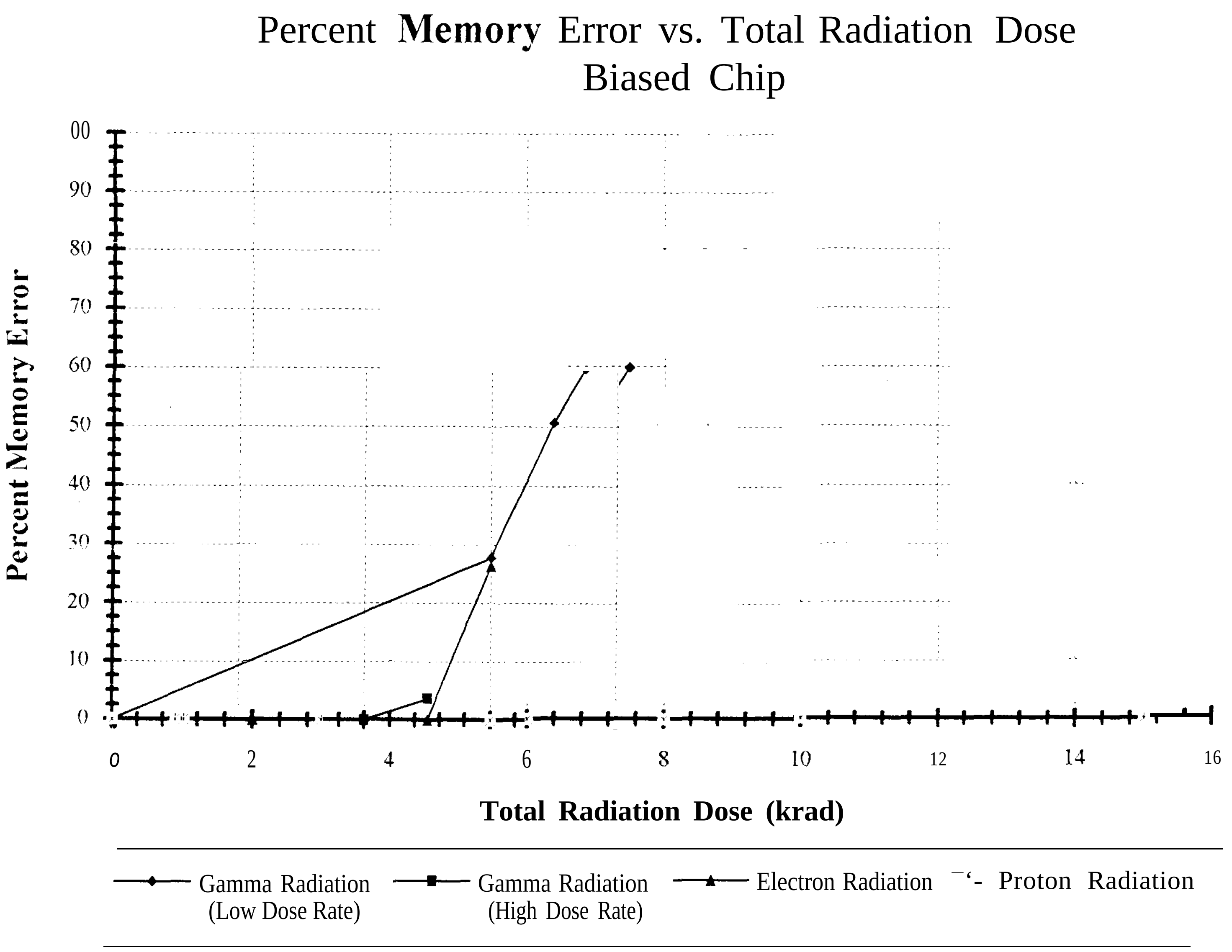

Figure 5.2a Percent Memory Error vs. Total Radiation Dose for All Biased Chips (Gamma, Electron, and Proton Radiations)

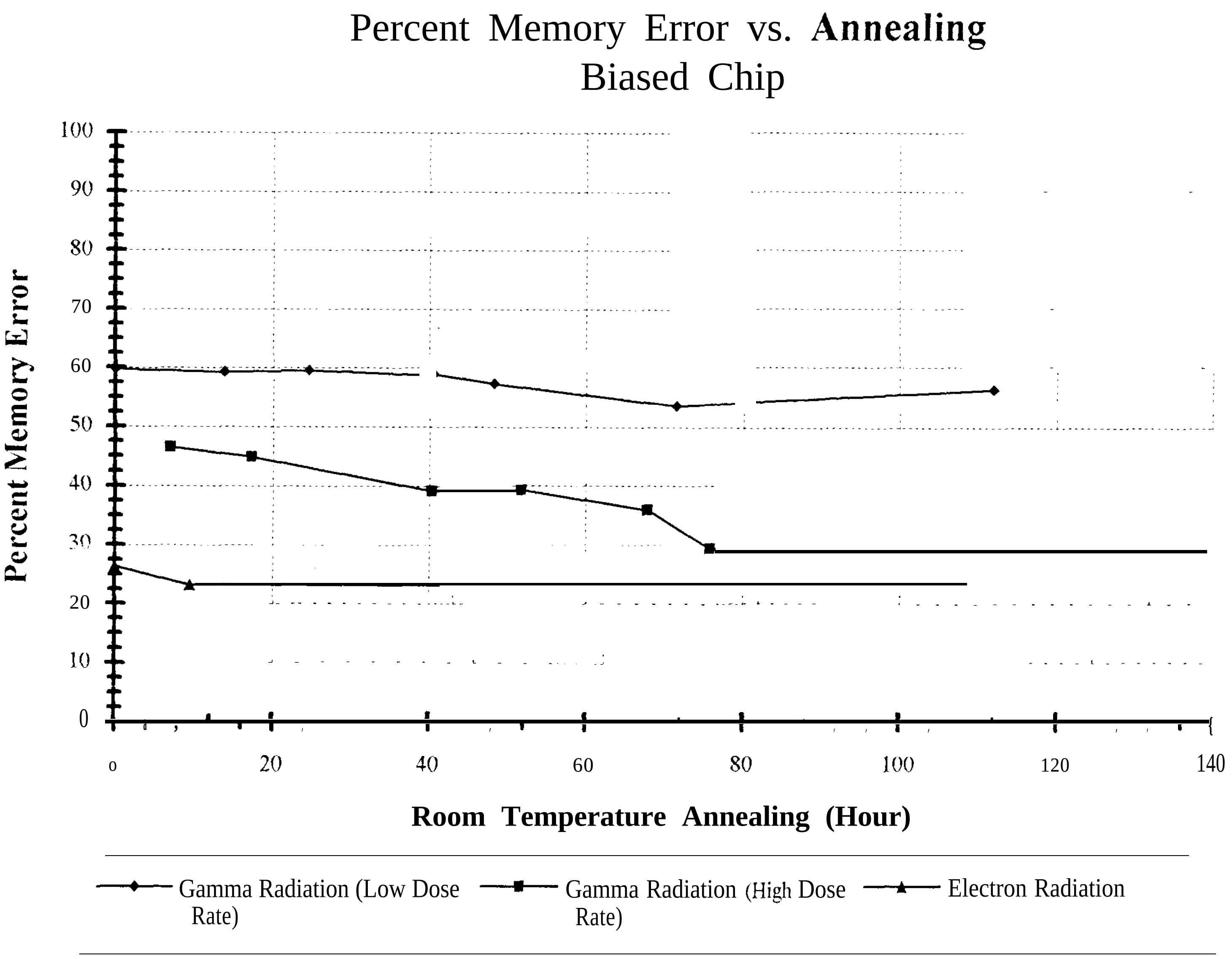

Figure 5.2b Percent Memory Error vs. Annealing for All Biased Chips (Gamma, Electron, and Proton Radiations)

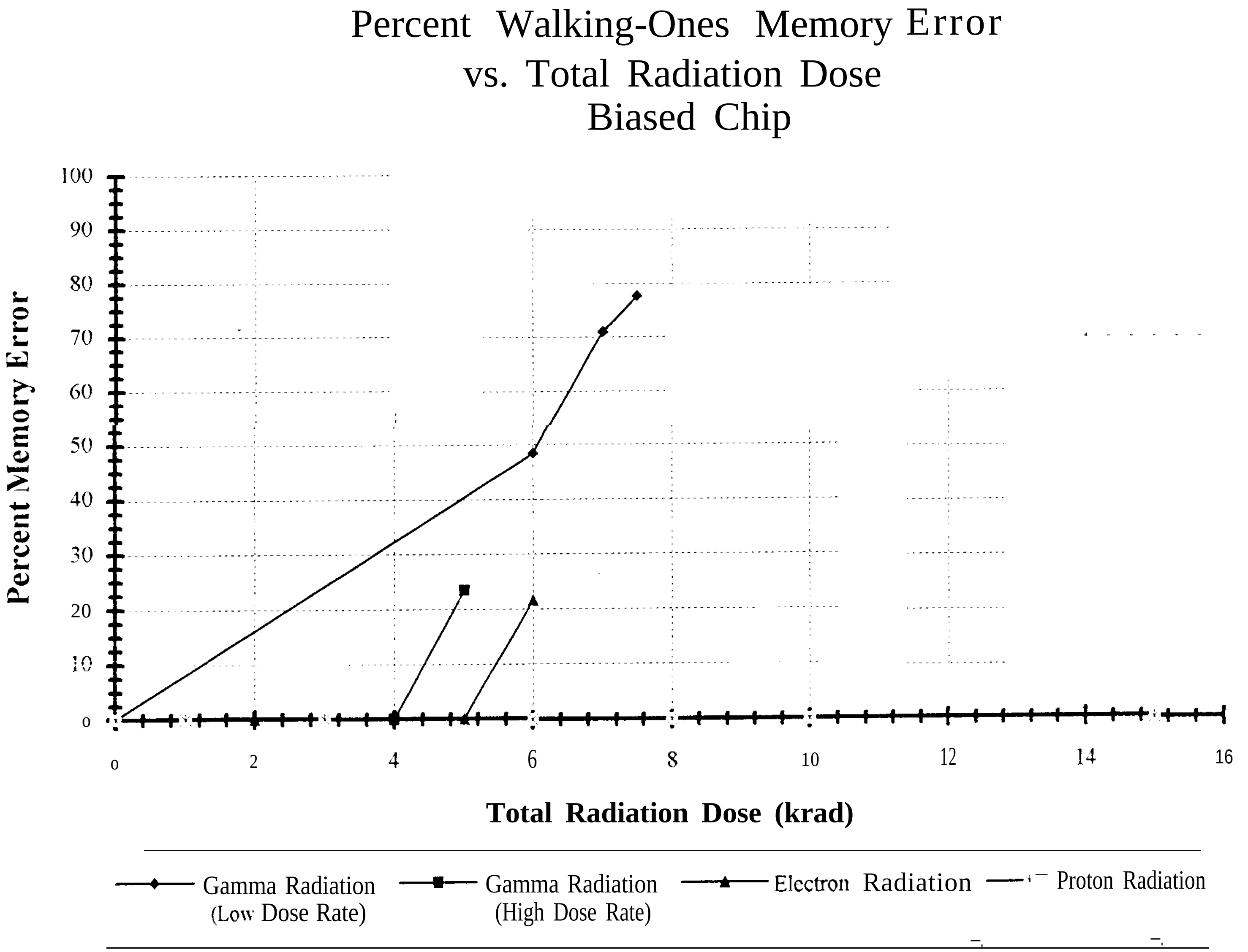

Figure 5.3a Percent Walking-Ones Memory Error vs. Total Radiation Dose for All Biased Chips (Gamma, Electron, and Proton Radiations)

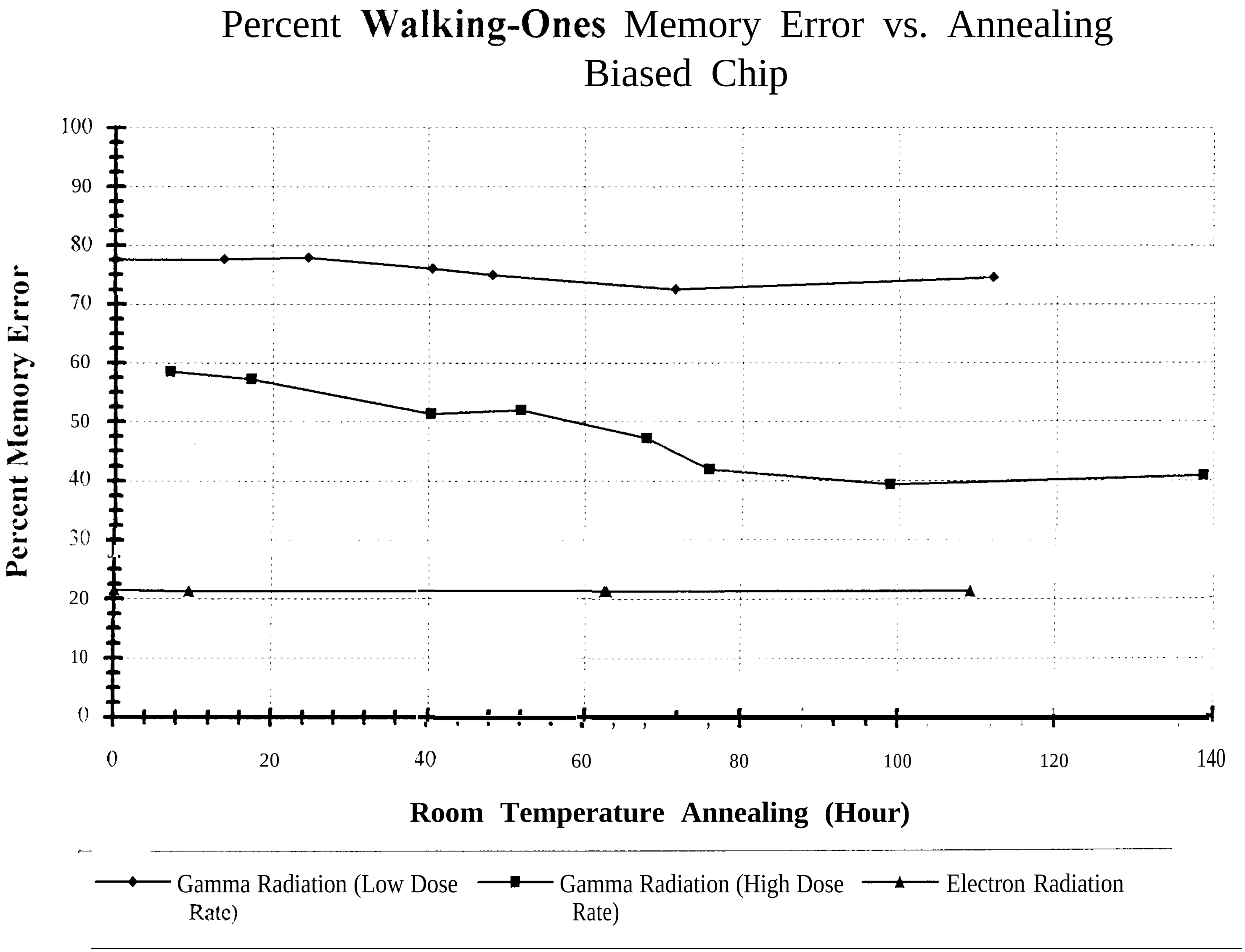

Figure 5.3b Percent Walking-Ones Memory Error vs. Annealing for All Biased Chips (Gamma, Electron, and Proton Radiations)

532 Synapse-Neuron Characteristics - Dose OOOk

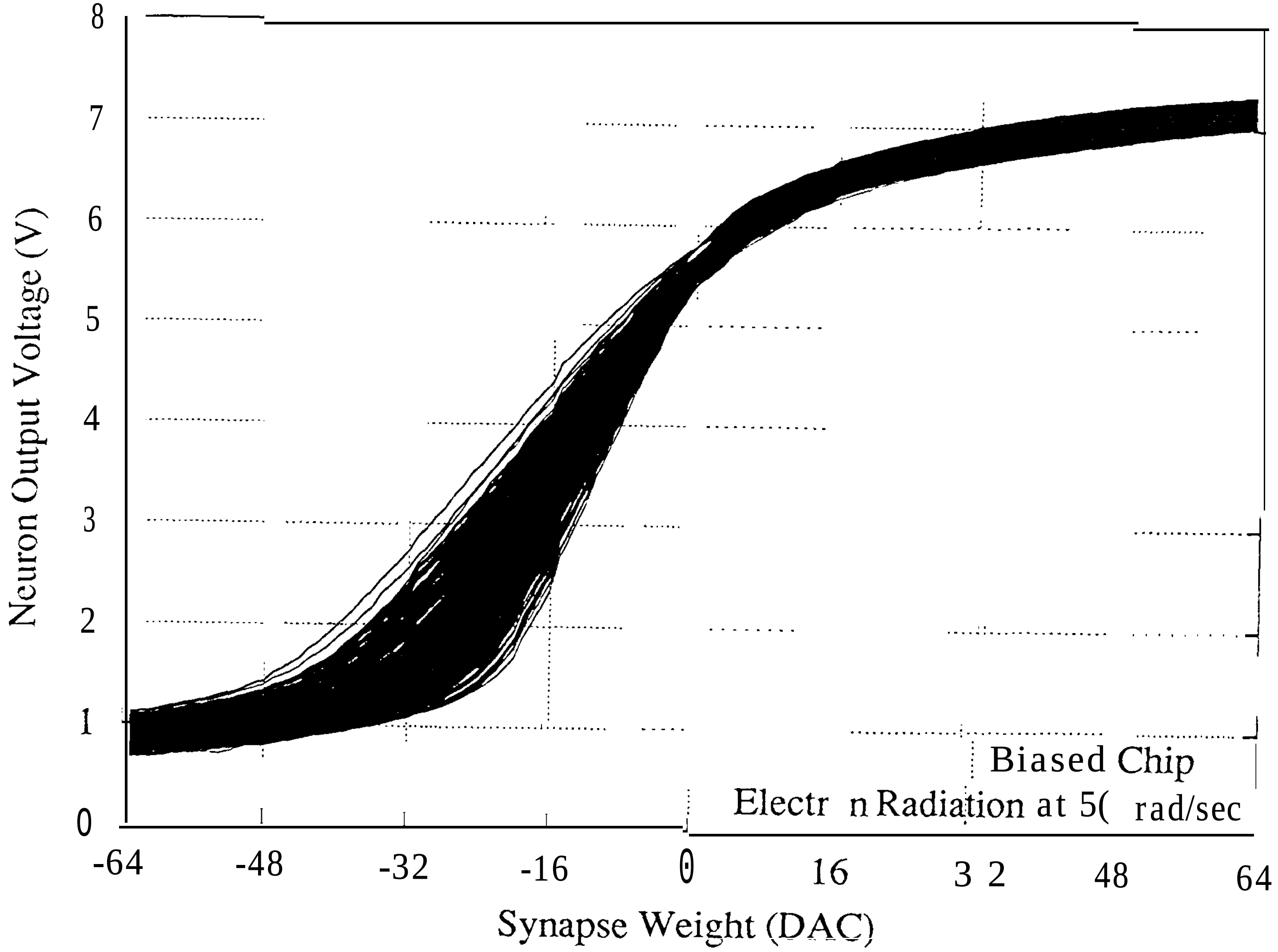

Figure 6.1a 532 Synapse-Neuron Characteristics at 0 krads
Biased Chip under Electron Radiation

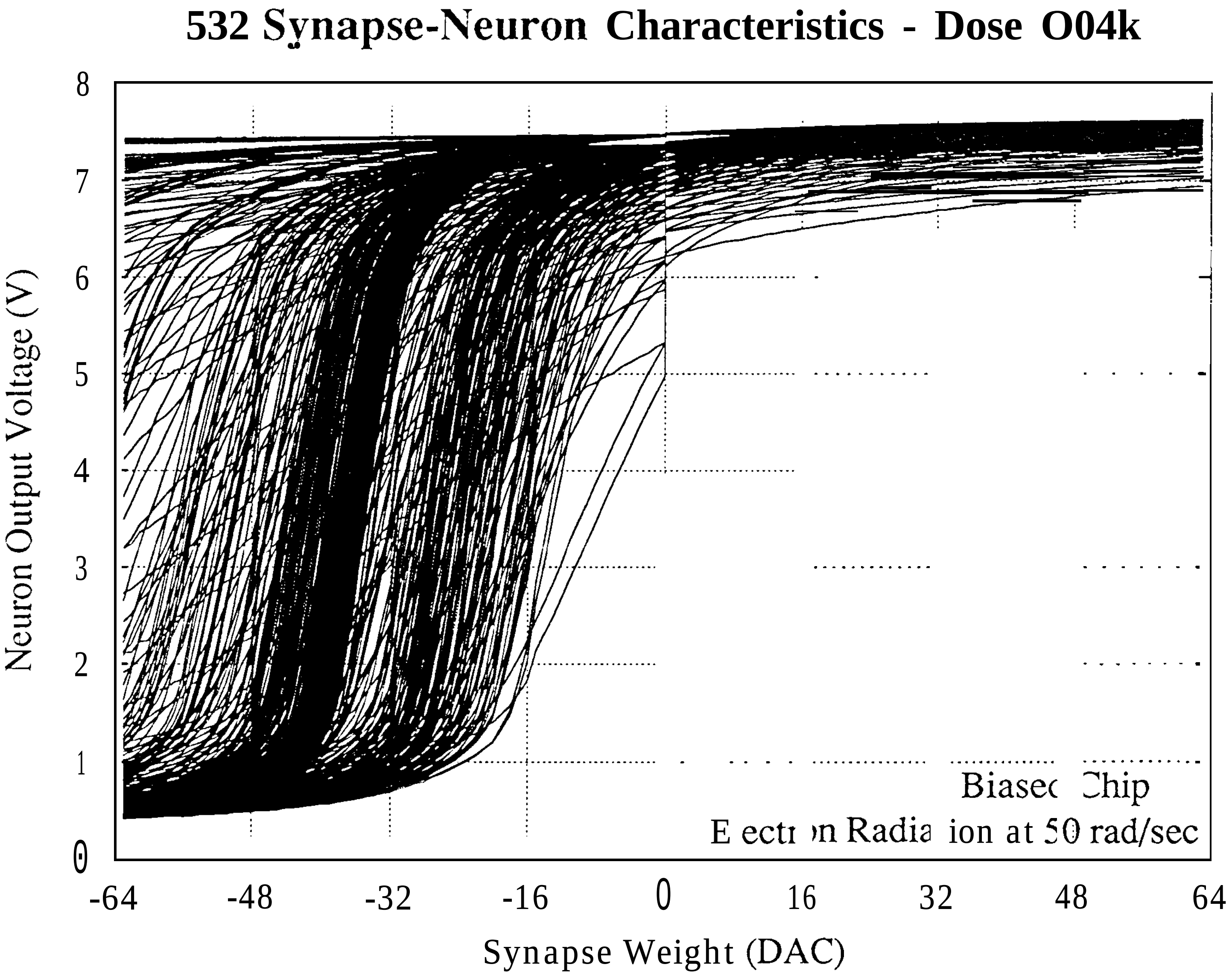

Figure 6.1b 532 Synapse-Neuron Characteristics at 4 krads Biased Chip under Electron Radiation

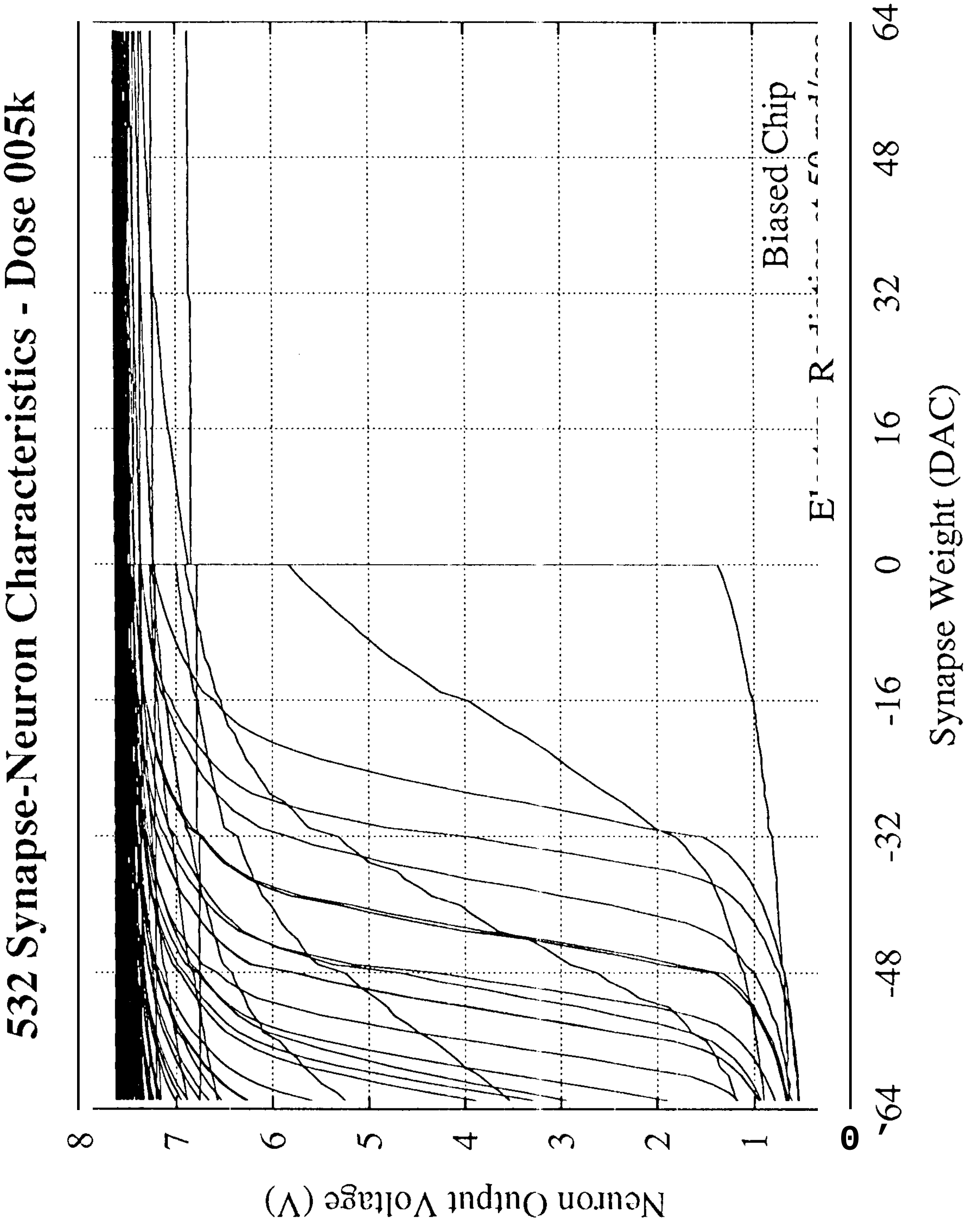

Figure 6.1c 532 Synapse-Neuron Characteristics at 5 krads Biased Chip under Electron Radiation

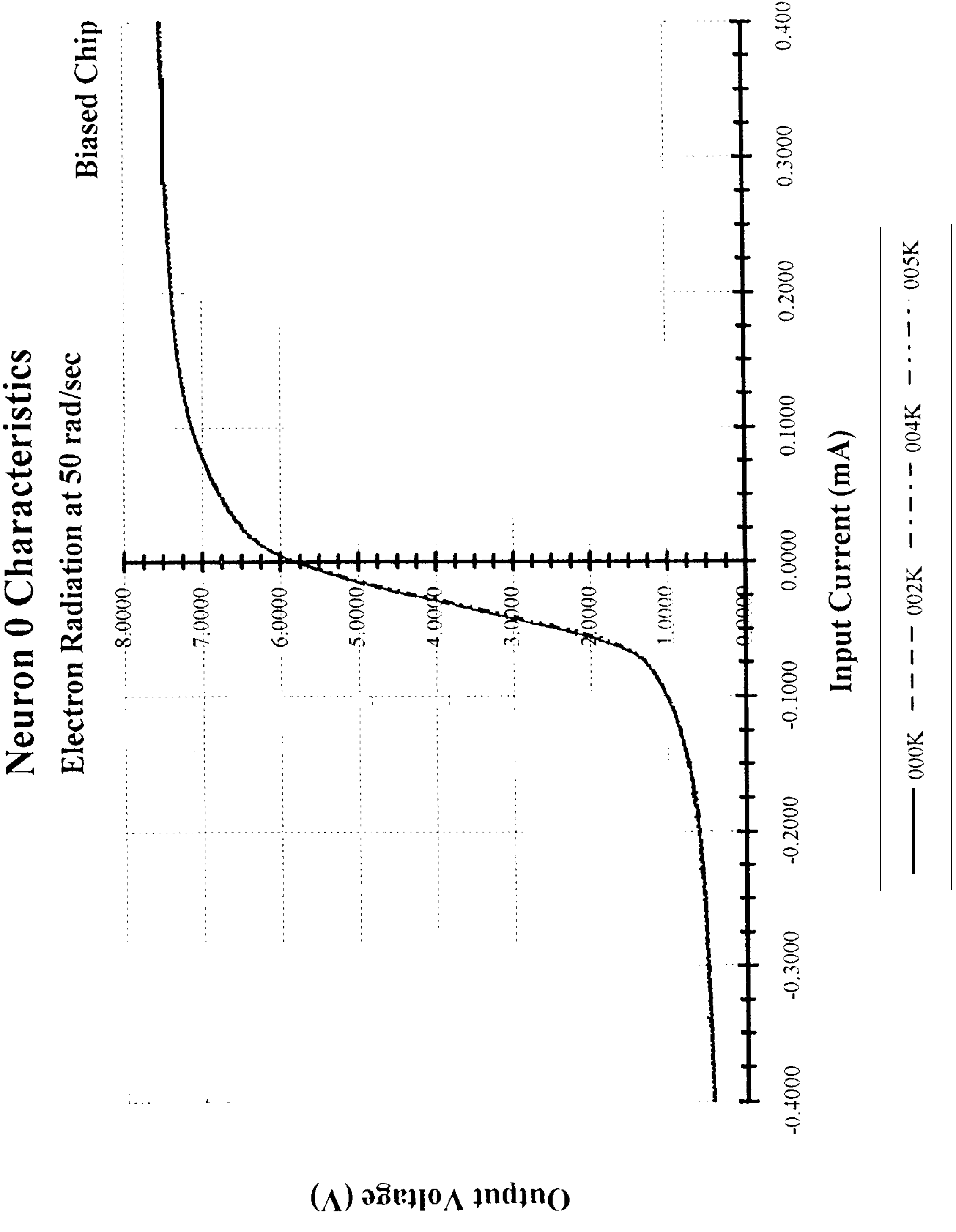

Figure 6.2 Neuron 0 Characteristics
Biased Chip under Electron Radiation

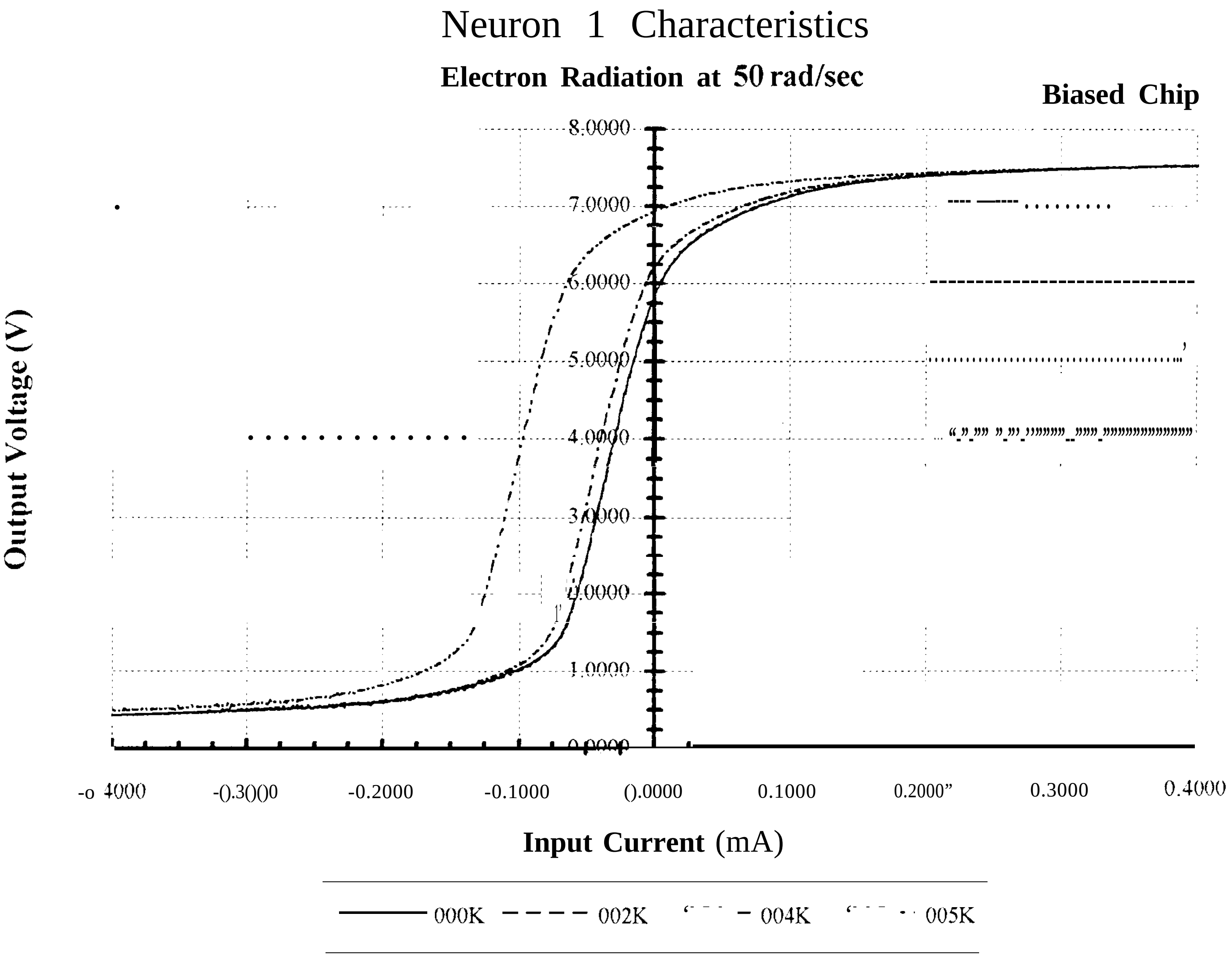

Figure 6.3 Neuron 1 Characteristics
Biased Chip under Electron Radiation

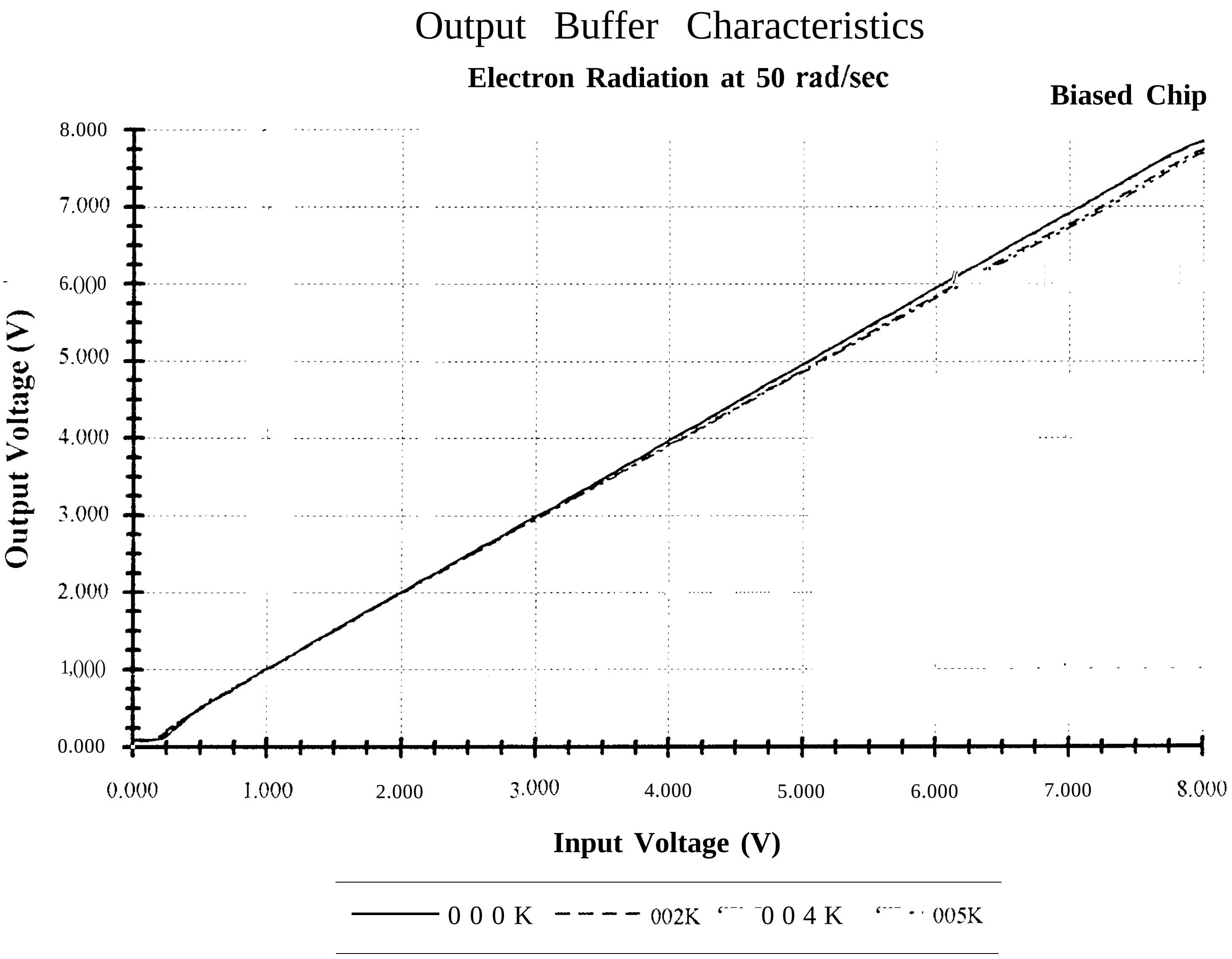

Figure 6.4 Ouput Buffer Characteristics
Biased Chip under Electron Radiation

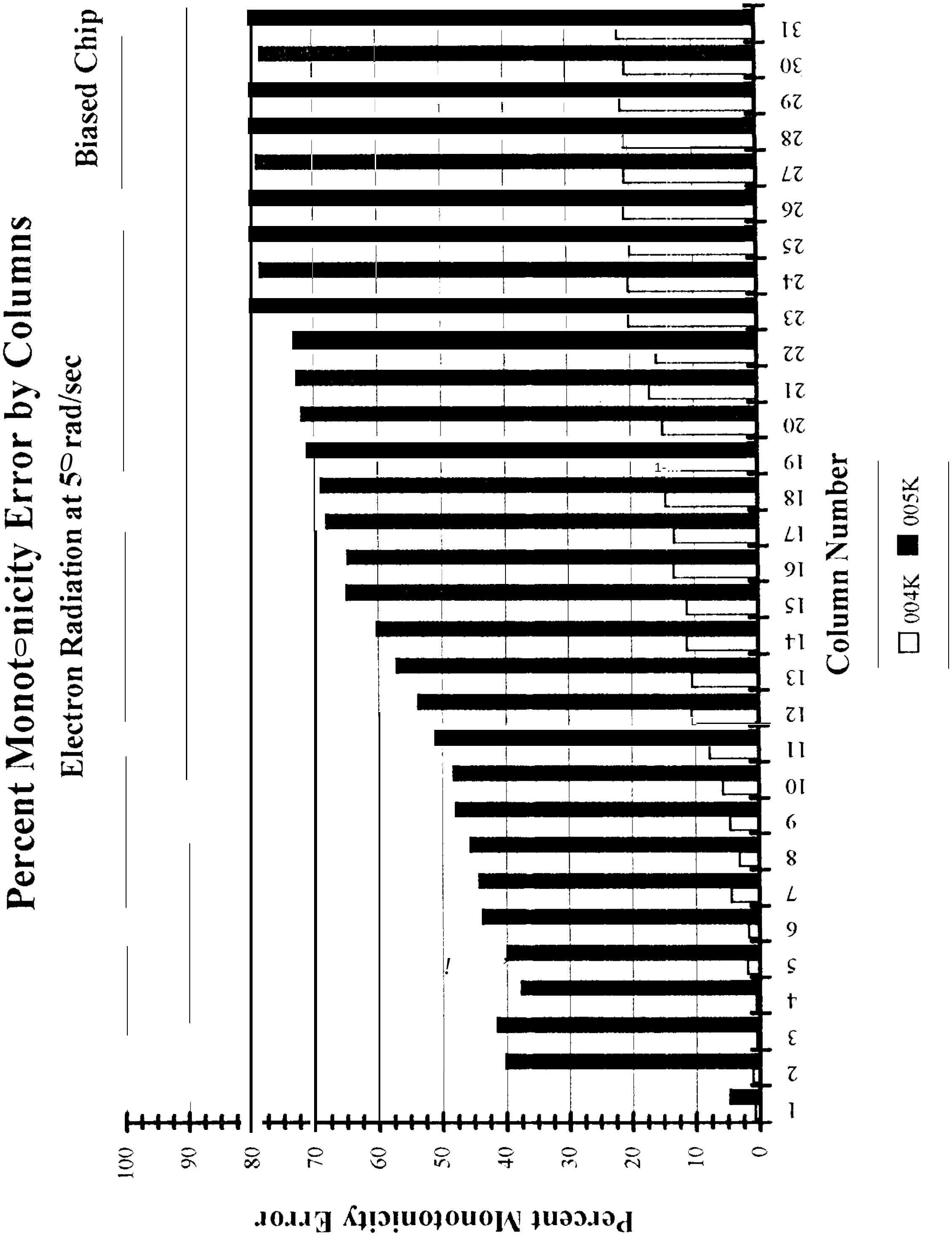

Figure 6.5 Percent Monotonicity Error by Columns Biased Chip under Electron Radiation

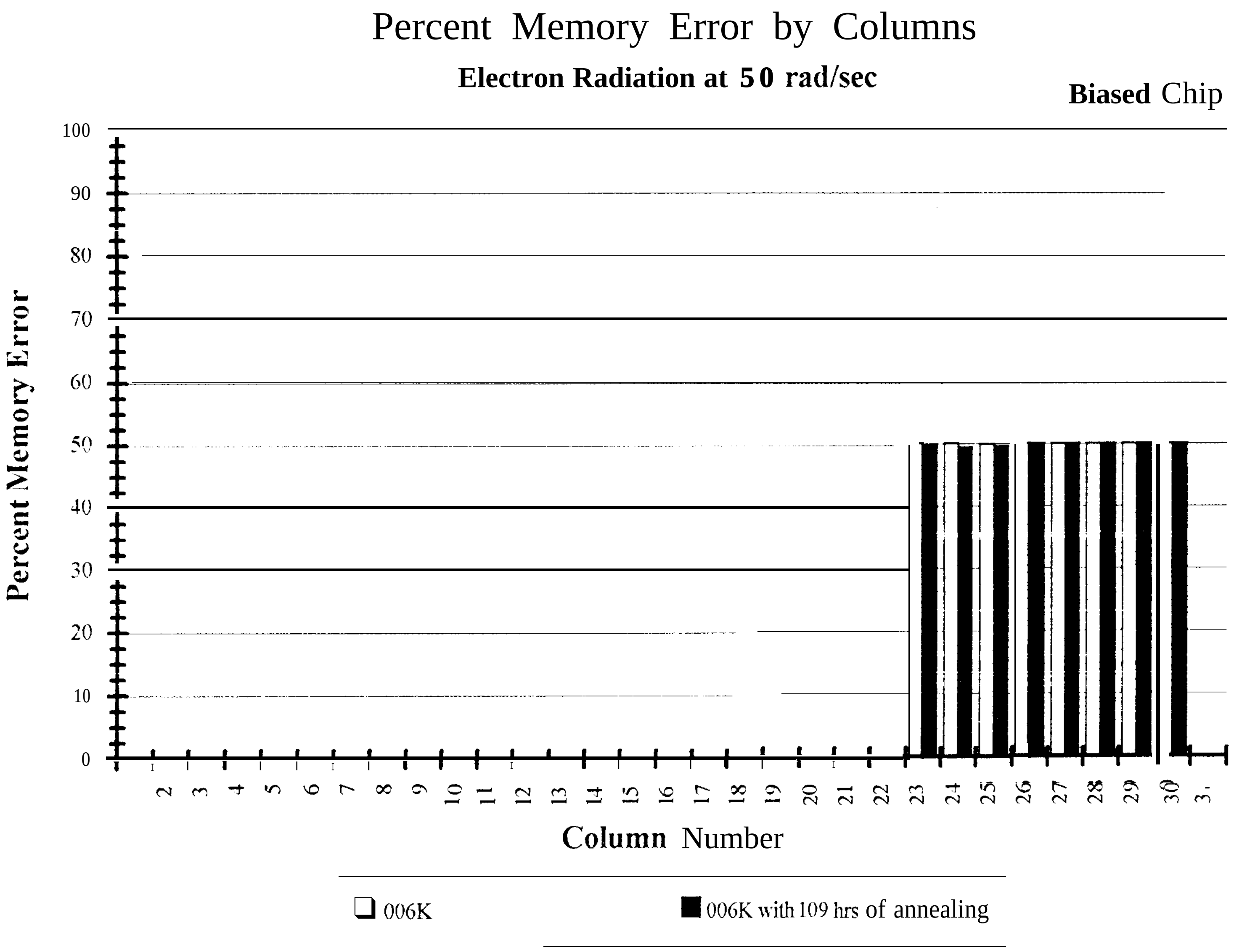

Figure 6.6 Percent Memory Errors by Columns Biased Chip under Electron Radiation

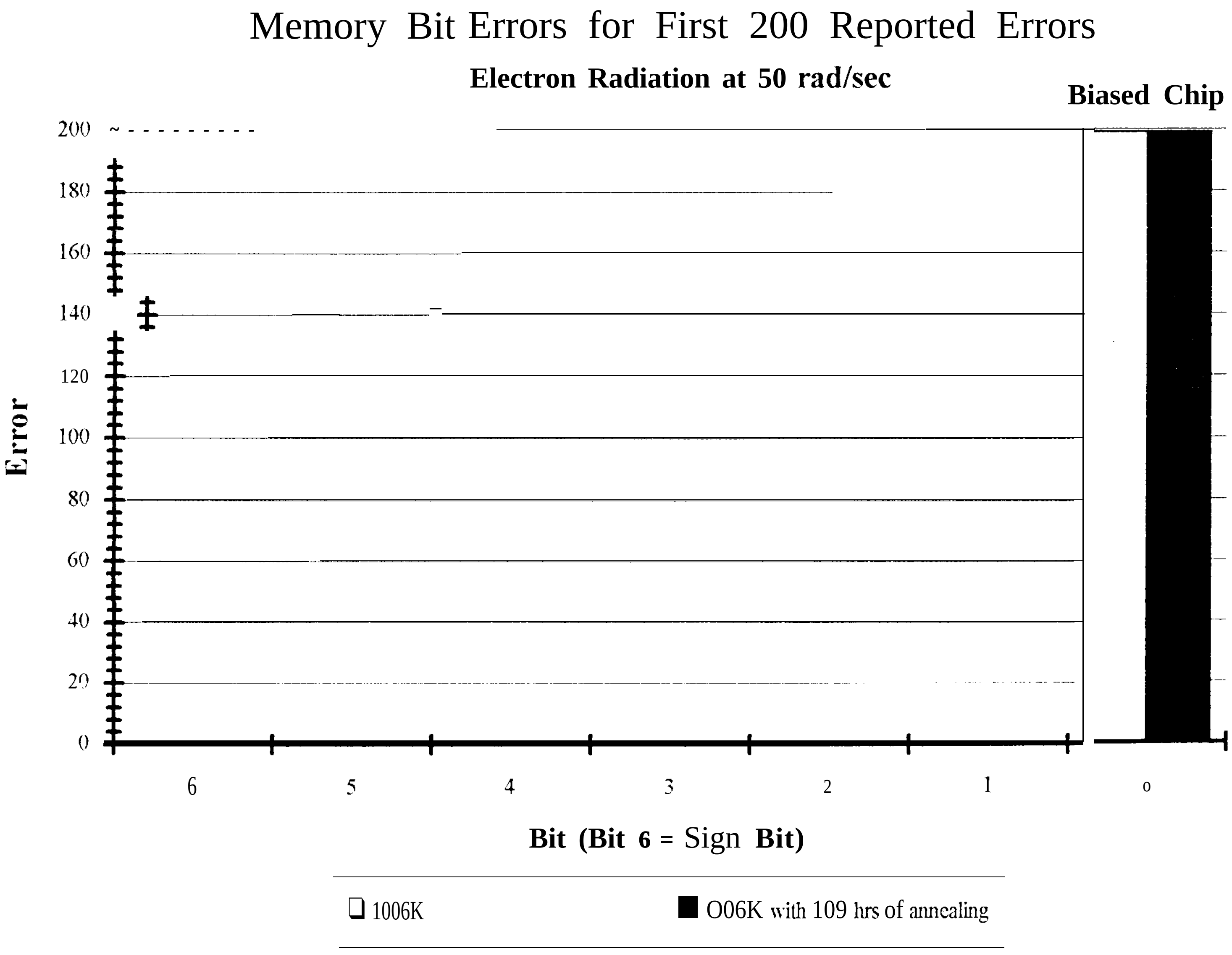

Figure 6.7 Memory Bit Errors for First 200 Reported Errors Biased Chip under Electron Radiation

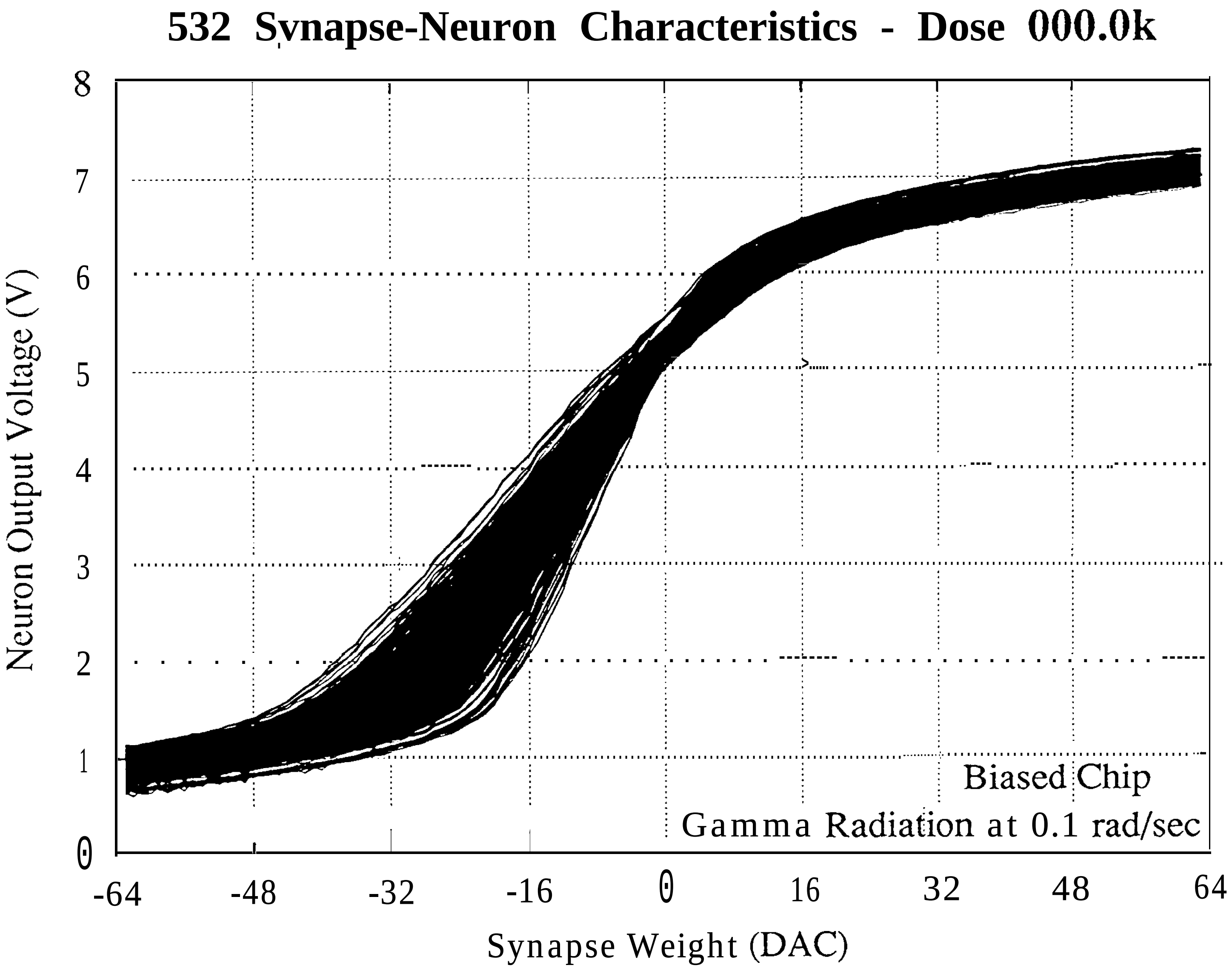

Figure 7.1a 532 Synapse-Neuron Characteristics at 0.0 krads Biased Chip under Gamma Radiation (Low Dose Rate)

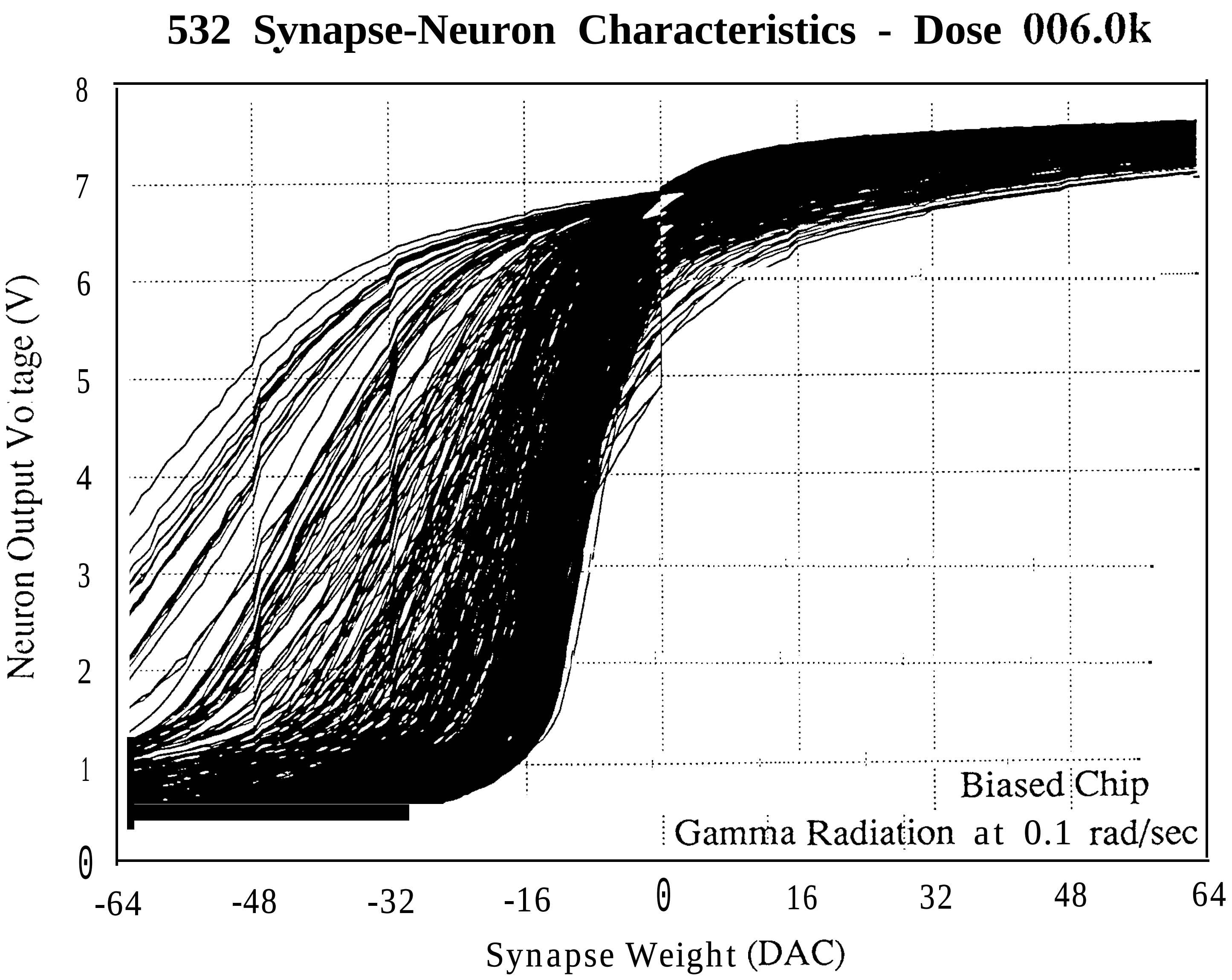

Figure 7.1b 532 Synapse-Neuron Characteristics at 6.0 krads Biased Chip under Gamma Radiation (Low Dose Rate)

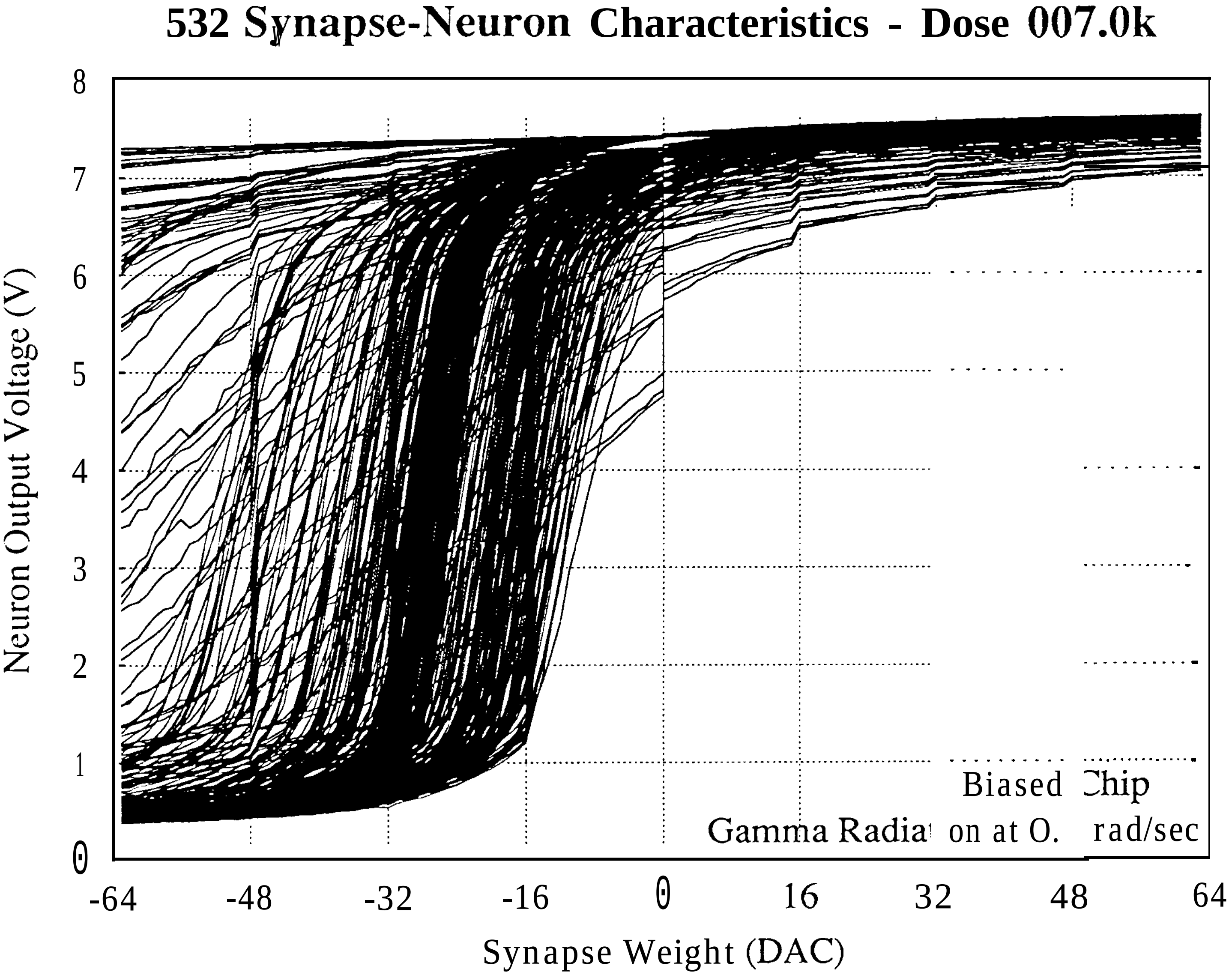

Figure 7.1c 532 Synapse-Neuron Characteristics at 7.0 krads Biased Chip under Gamma Radiation (Low Dose Rate)

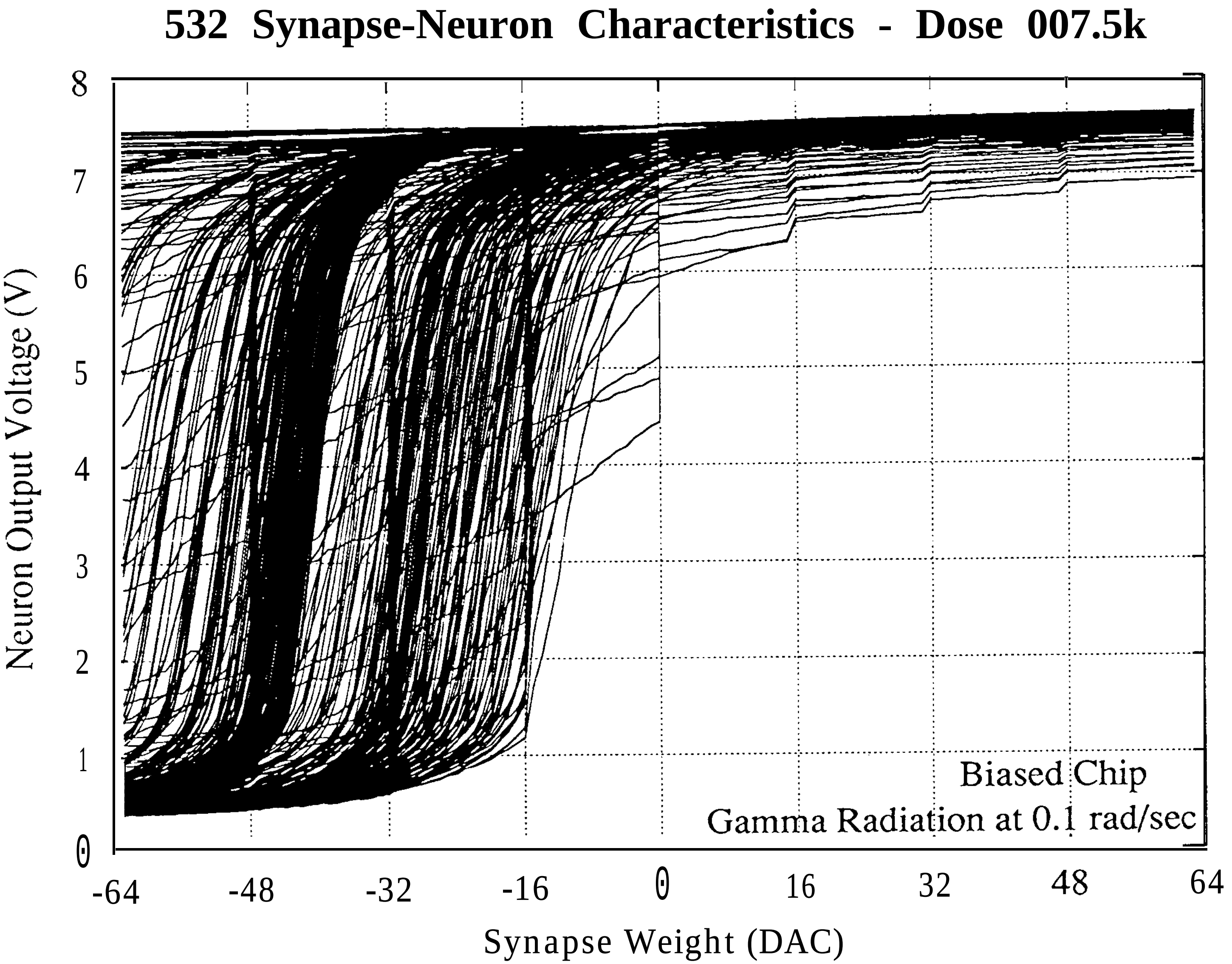

Figure 7.1d 532 Synapse-Neuron Characterisics at 7.5 krads Biased Chip under Gamma Radiation (Low Dose Rate)

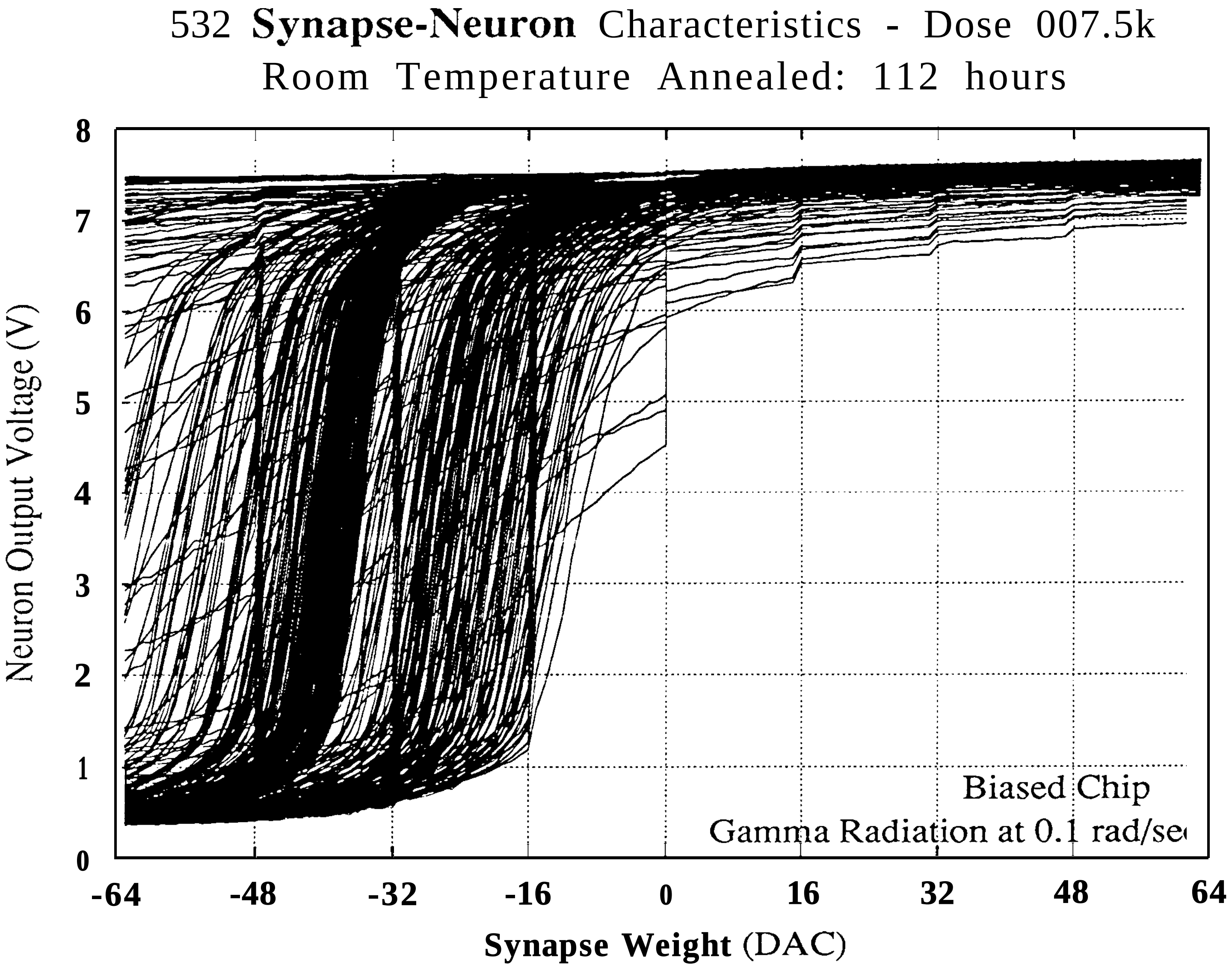

Figure 7.1e 532 Synapse-Neuron Characteristics at 7.5 krads, Annealed 112 hours Biased Chip under Gamma Radiation (Low Dose Rate)

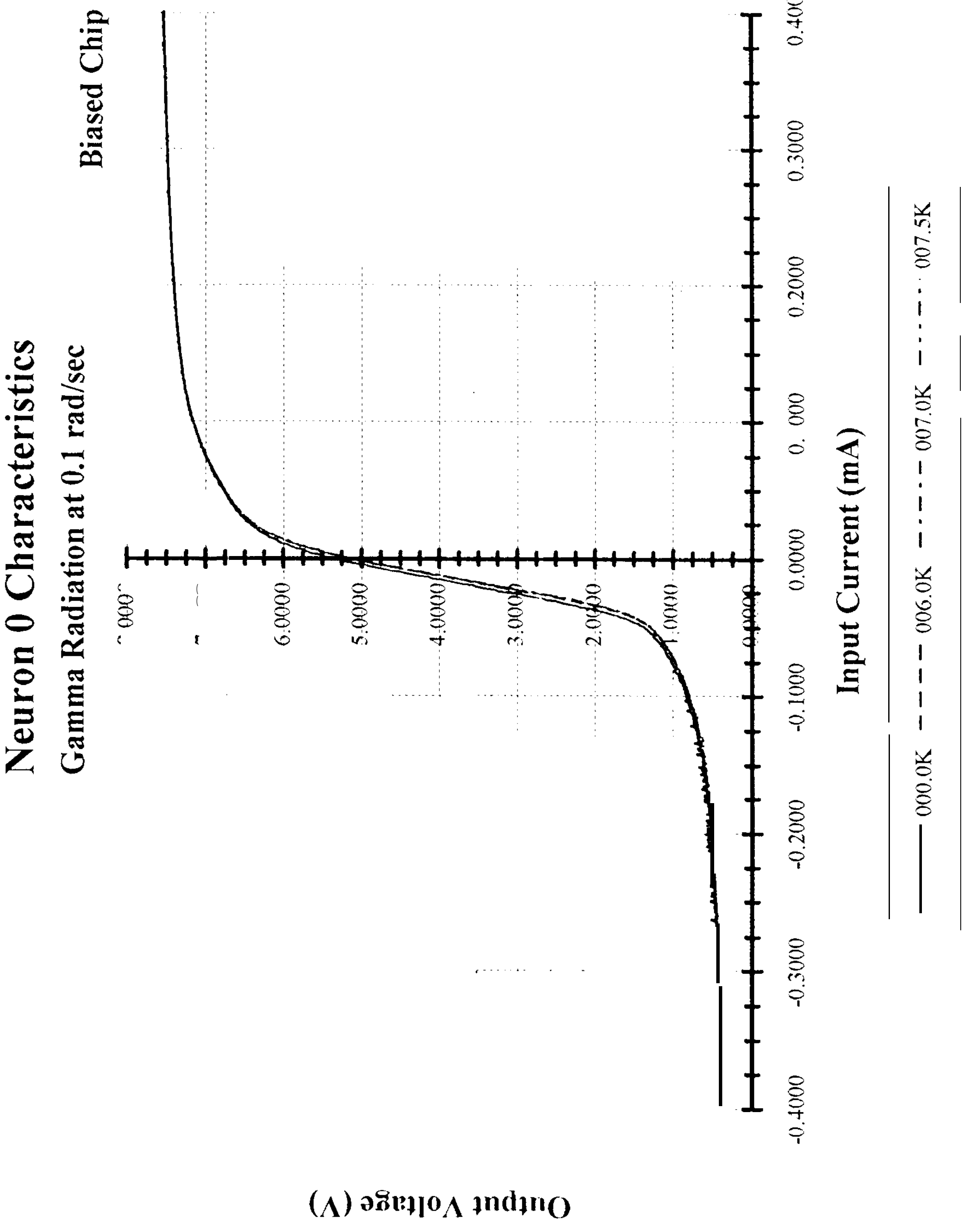

Figure 7.2 Neuron 0 Characteristics
Biased Chip under Gamma Radiation (Low Dose Rate)

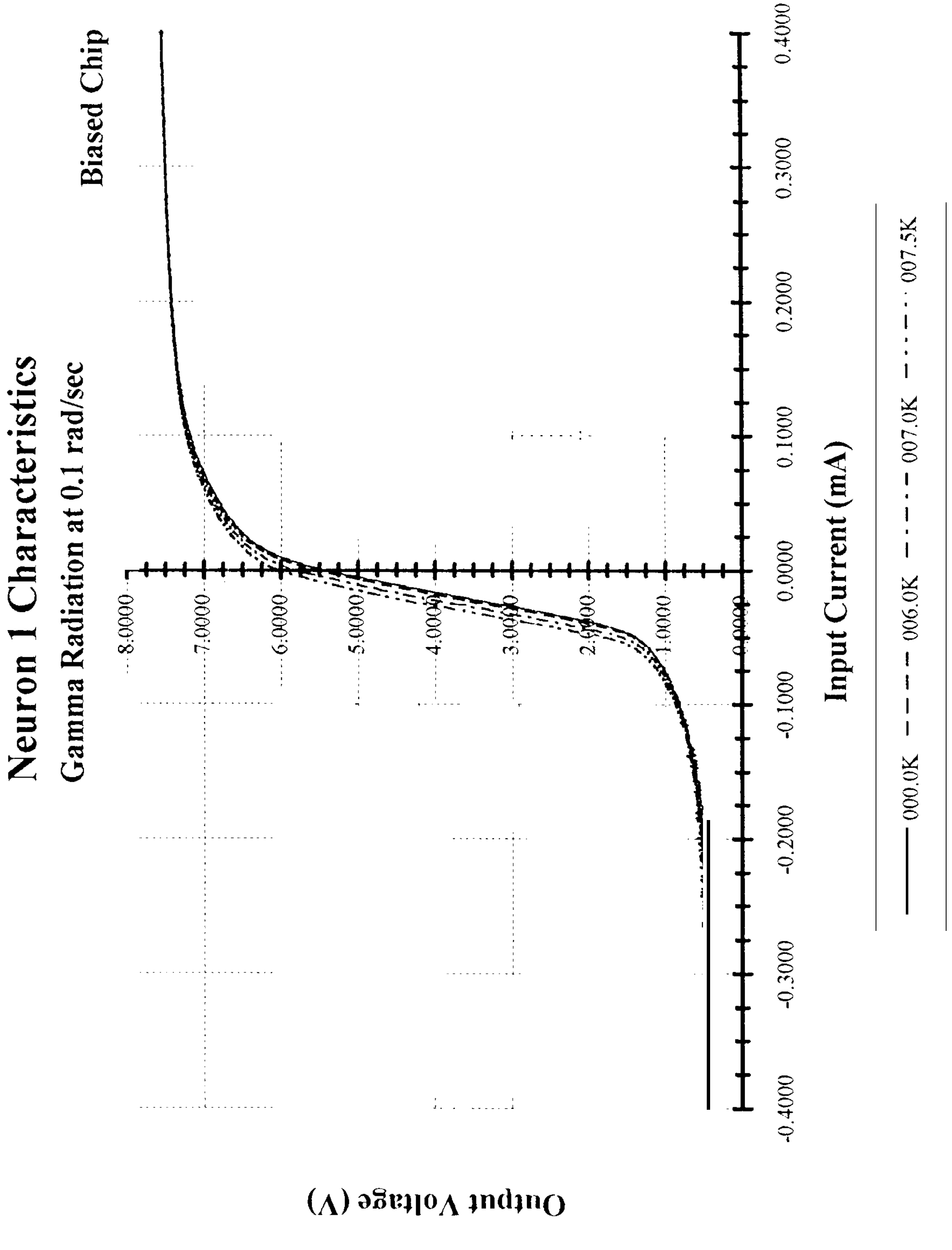

Figure 7.3 Neuron 1 Characteristics
Biased Chip under Gamma Radiation (Low Dose Rate)

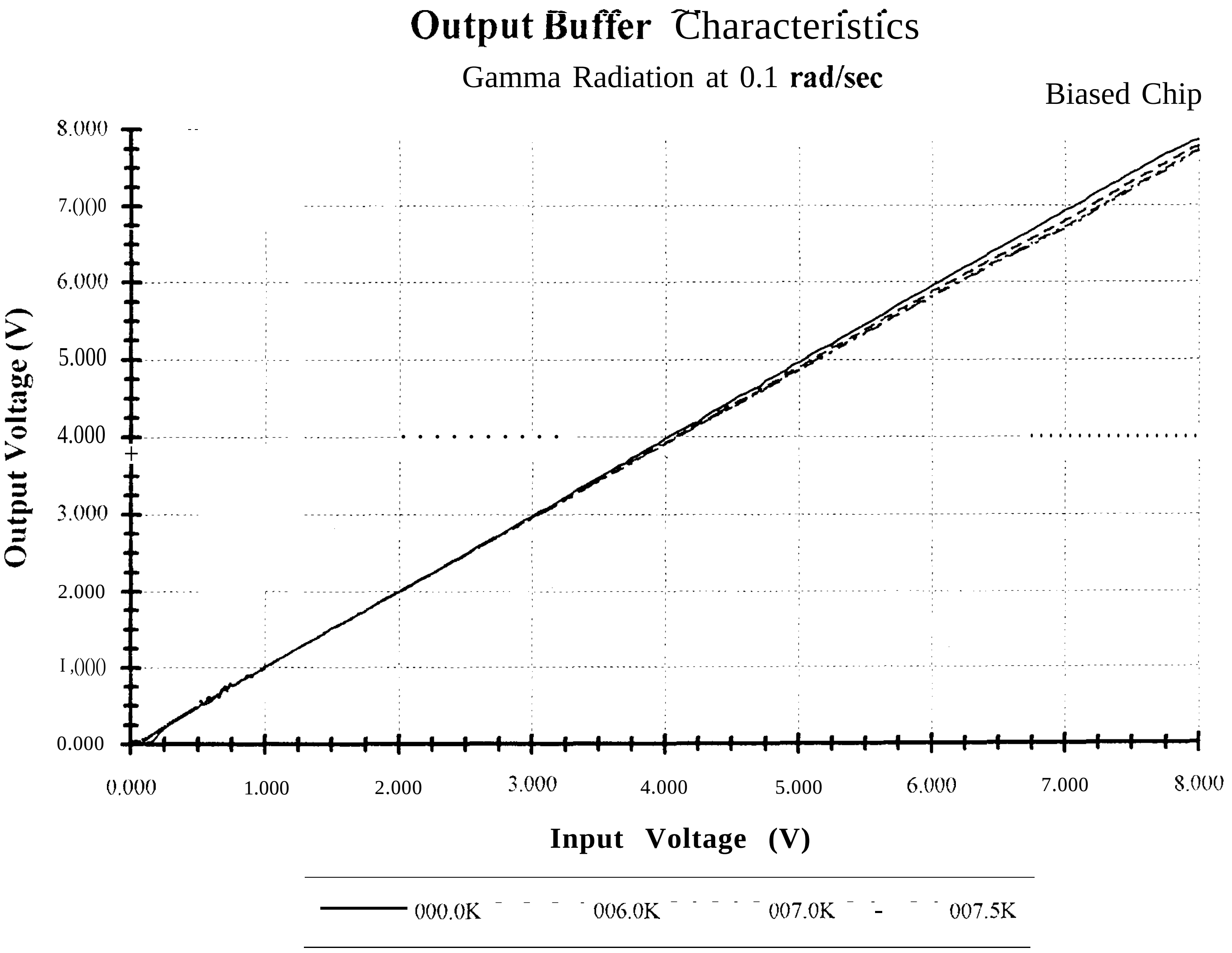

Figure 7.4 Output Buffer Characteristics
Biased Chip under Gamma Radiation (Low Dose Rate)

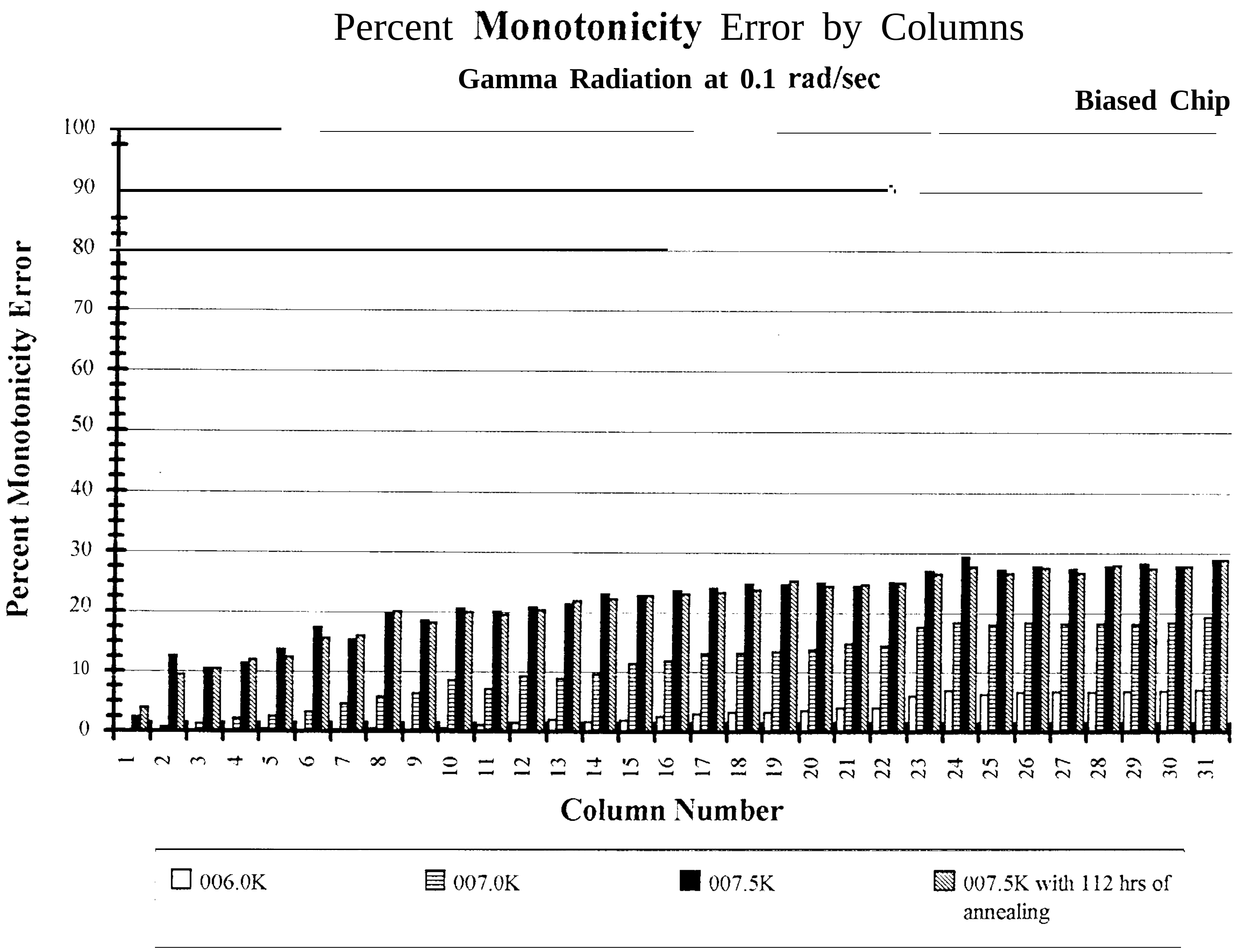

Figure 7.5 Percent Monotonicity Error by Columns
Biased Chip under Gamma Radiation (Low Dose Rate)

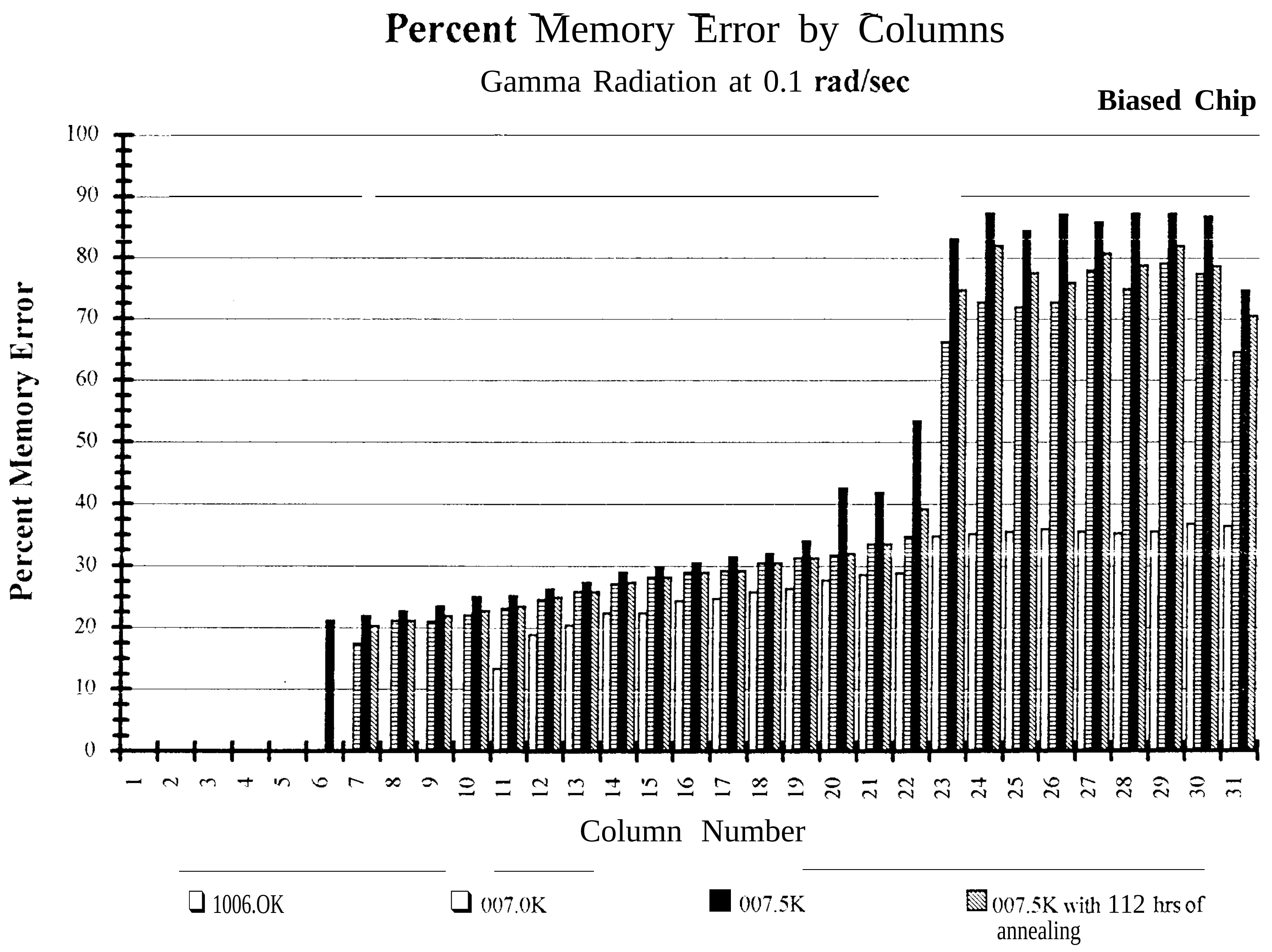

Figure 7.6 Percent Memory Errors by Columns
Biased Chip under Gamma Radiation (Low Dose Rate)

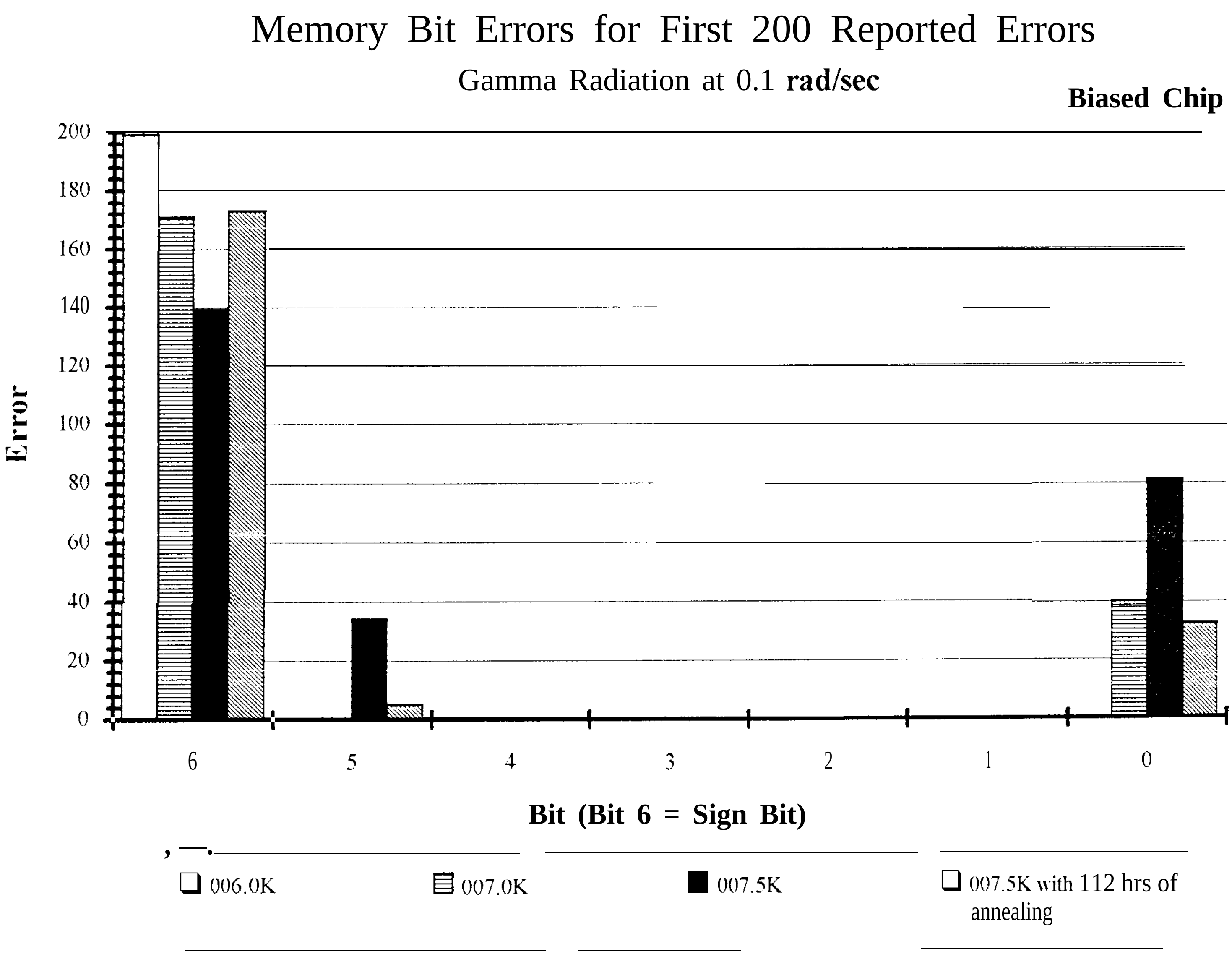

Figure 7.7 Memory Bit Errors for First 200 Reported Errors Biased Chip under Gamma Radiation (Low Dose Rate

Chapter 5

Discussion

MOS transistors were originally thought to be radiation-hard because their transfer characteristics did not depend On minority carrier lifetime [1 1]. (Minority carrier lifetime is the most radiation-sensitive material parameter and the main cause of radiation-induced failure in bipolar transistors.) However, this optimism was proven to be ill founded. Both n-FET and p-FET devices were found to be very sensitive to ionizing radiation resulting inlarge changes in threshold voltage and transconductance.

ionizing radiation causes detrimental effects on the characteristics of FET devices and circuits. The threshold voltages, current driving capabilities, and leakage currents of transistors change as a function of a number of factors: the total (lose of radiation received and its energy; the bias voltages applied during the irradiation; the geometry, type, and method of fabrication of the transistor; the dose rate at which the radiation is delivered; the temperature during the irradiation; the bias, time, and temperature after the irradiation is completed [111. Changes in the properties of the devices can lead to significant changes in the characteristics of the integrated circuits of which they arc the primary elements.

The response of integrated circuits may be understood in terms of the combined response of the individual transistors [11]. 1 Iowever, in complex circuits such as the artificial neural network chip, the analysis can be difficult because of the large number of possible bias configurations and circuit paths.

The results of the 7 different neural network chips as described in the previous chapter will be collated here, analyzing the differences between the biased and unbiased chips, the effects from the three different radiation sources, and comparing the two dose rates used on the biased chips under gamma radiation. Also, the degradation and annealing behavior of the chips will be examined in more detail.

5.1 Biased vs. Unbiased

In addition to the biased chips degrading more severely with less radiation as compared to the unbiased chips, there were other differences observed. For the unbiased chips, there were radiation effects to the neuron circuit as observed in the neuron () and neuron 1 tests. The sigmoidal curve of the neuron became steeper as the radiation dose was increased, while the unbiased chips showed no changes. Both chips did, however, show the light left shift of the sigmoidal curve in the neuron 1 test.

Also, there were differences in the performance of the output buffer in the unbiased and biased chips. While the biased chips did not show any degradation, the unbiased chips showed a “kink” where the output voltage was not a linear function of the input voltage.

Finally, the annealing behavior of the chips after the final cumulative dose was different. The unbiased chips showed partial recovery with room temperature annealing while the biased chips did not show any significant recovery. For example, the synapse memory for the unbiased chips in the electron and gamma radiations recovered fully. The synapse-neuron curves also showed recovery for those unbiased chips as some of the damaged sigmoidal shaped curves shifted back to the right. For the biased chips, there was only a slight change in the synapse-neuro n curves and in the memory errors.

Ionizing radiation causes significant changes in the. characteristics of FET devices due to surface effects [11]. The major mechanism of degradation is due to the creation of oxide-trapped charge caused by radiation-induced positive charge buildup in the gate-oxide region. Ionizing radiation also causes surface states (or interface traps) at the Si/SiO_2 interface. Oxide-trapped charge causes a negative shift in transistor threshold voltages; interface traps cause a decreased subthreshold slope in transistors. The electric field applied across the oxide in a FET device during irradiation has a dominant effect on the radiation damage introd uced. Positive fields (positive bias voltage applied to the gate electrode) cause the vorst-case damage.

Threshold voltage shifts for both n-FET and p-FET devices thus depend upon the bias applied during irradiation: the voltage applied to the gate electrode has a first-order effect, and the voltage applied to the source and drain can also be significant [11]. The p-FET threshold typically shifts monotonically negative as the total dose is increased. The n-FET threshold response can be more complicated, shifting in the negative direction initially and, as the total dose increases, eventually turning around and shifting in the positive direction as the compensation of oxide-trapped charge by interface trap charge becomes more important.

The part of an MOS structure most sensitive to ionizing radiation is the oxide insulating layer. When ionizing radiation passes through the oxide, the energy deposited creates electron/hole pairs by breaking silicon-oxygen bonds [11]. Some of the radiation-induced charge carriers recombine, whereas most of them drift in the applied electric field, toward the appropriate electrode (gate or silicon substrate).

Because of their higher mobility, electrons rapidly drift toward the gate which is the positive electrode and flow into the external circuit [11]. Since thermally grown oxides have low concentrations of electron traps, nearly all the electrons exit the oxide region.

The holes that escape initial recombination are relatively immobile and remain behind near their point of generation, causing negative voltage shifts in the electrical characteristics of FET devices [11]. However, over a period of time, the holes undergo a rather anomalous stochastic hopping transport through the oxide in response to any electric fields present. This hole transport process, which is very dispersive in time, gives rise to a short-term, transient recovery in the voltage shift. It is sensitive to many variables including primarily applied field, temperature, and oxide thickness. upon reaching the Si/SiO_2 interface, some of them are captured in long-term trapping sites (the hole trap distribution usually extends a few nanometers from the Si/SiO_2 interface), and cause a remnant negative voltage shift that is not sensitive to the silicon surface potential and which can persist in time from hours to years. The holes trapped at this interface have a much larger effect on the voltage shifts than those trapped at the gate electrode interface. This long-lived radiation-induced voltage shift component is the most commonly observed form of radiation damage and is very sensitive to the electric field and temperature.

The oxide trapped charges anneal (recombine) on a linear basis with a logarithmic time scale [11]. The annealing slopes are linearly dependent on absolute temperature. A tunneling process has been hypothesized to be responsible for long-term annealing, as well as for the eventual saturation magnitude for oxide charges.

When silicon is thermally oxidized, the interface between the amorphous oxide and the crystalline silicon is generally deficient in oxygen (or abundant in silicon), giving rise to strained as well as uncompleted, or “dangling,” silicon bonds [11]. These dangling bonds act as interface traps with energy levels within the forbidden bandgap at the SiO_2/Si interface. Before irradiation, the areal density of interface traps in a processed FET device is in the range of 109-101°” traps/cm^2 which normally are not much of a problem. When the FET devices are exposed to ionizing radiation, additional interface traps can be generated at the SiO_2/Si interface, resulting in discernible and often detrimental effects in devices. In addition, other types of stress, including avalanche electron or hole injection and high-field stressing, are also known to create interface traps at the oxide-semiconductor interface. Annealing Of interface traps does not occur at normal operating temperatures except with very high densities of interface traps.

5.2 Effects of Different Radiation Sources

The three types of radiation, electron, gamma, and proton, showed different effects and degradation on the chips whether unbiased or biased. A summary of these differences will be stated, first for the unbiased chips and then for the biased hips.

5.2.1 Unbiased Chips

Electron radiation for the unbiased chip was the first to cause any damage and can be observed in the monotonicity test where errors started occurring at 40 krads (Figure 3.1a in Chapter 4) and in the memory test where errors started occurring at 120 krads (Figure 3.2a and 3.3a). The "kink" in the output buffer was observed at a total dose of 30 krads, and the shifting of the sigmoidal synapse-neuron curves was observed at 60 krads.

The unbiased chip having undergone gamma radiation was next to demonstrate any damage. This can be observed in the analog characteristics where the "kink" at 40 krads developed in the output buffer curve, monotonicity errors developed at 50 krads, and at 70 krads the synapse-neuron curves started shifting to the left. Memory errors started developing at 190 krads. Similarly, the proton radiation for the unbiased chip also showed the onset of damage **on** the synapse memory at about the same time.

However, there *were no* memory errors *at 150* krads, but at 200” krads there were more memory errors for the proton radiation than for the gamma radiation, The analog portion of the neuron circuit for the proton radiation had effects starting at 80 krads. At this dose, the “kink” developed in the output buffer and monotonicity errors developed in the synapse-neuron curves which also started shifting at 150 krads. The proton radiation seemed to cause more degradation in the hardware than the other two types of rad i at ion. For example, at 200” krads, there were comparatively more monotonicity and memory errors.

The degradation from electron and gamma radiations seemed to saturate (flatten out) after the unbiased chips received a certain cumulative dose of radiation. Monotonicity errors for the electron radiation saturated earlier at 120 krads with a higher error of 73% than for the gamma radi ation at 170 krads with a lower error of 62%. Memory errors saturated at the same 30 % for the standard test and 48% for the walking-ones test for both electron and gamma radiations. Similar to the monotonicity errors, the onset of saturation was earlier for the electron radiation than for the gamma radiation. Proton radiation damage seemed to saturate at a higher cumulative dose with a larger error of 90% for the monotonicity and memory tests. The onset was at 200 krads for the monotonicity errors and at 280 krads for the memory errors.

Post radiation annealing of the chips also showed differences among the three types of radiation. Both the electron and gamma irradiated chips showed annealing behavior, whereas for the proton radiation, the unbiased chips did not show any signs of annealing. This might be attributed to the fact that the chip was severely damaged due to the high total dose (440 krads). The chip with gamma radiation annealed faster than the electron radiation. The synapse memory fully recovered after 17 hours of annealing at room temperature for the gamma radiation but needed at least 100 hours for the chips that underwent electron radiation (1 'igures 3.2b and 3.3 b). This was also observed in the synapse-neuron curves and in the monotonicity test (Figure 3.1b). To summarize, the electron radiation was the first to cause any damage on the unbiased chips; the proton radiation was the last to cause any damage, but the damage was more severe; and the gamma radiation showed faster annealing.

5.2.2 Biased Chips

Similar radiation effects were observed for the biased chips exposed to electron and gamma rad i at i **ons.** For example, electron radiation would show analog failures such as monotonicity errors (Figure 5.1a in Chapter 4) and shifting of synapse-neuron curves 1 krad earlier than the gamma radiation, whereas the gamma radiation would show digital failures (memory errors) 1 krad earlier than the electron radiation (Figures *5.2a* **and** 5.3a). nut, both types of radiation started having effects around 4 krads and became nonfunction al around 7 krads. Annealing behavior was similar to the unbiased chips where gamma radiation annealed slightly faster and better than the electron radiation (Figures 5. lb, 5.2b, and 5.3 b).

The chips with proton radiation, however, showed significant difference in degradation. There were no radiation effects such as changes in the synapse-neuron curves, monotonicity errors, or memory errors for a cumulative dose of 15 kinds. Again, similar to the unbiased chip, the proton radiated chip did not show any annealing, since at 30 krads the biased chip might have suffered severe damage. Thus, the main difference among the three types of radiation was that the proton radiation had the least effect on the biased chip.

5.3 Dose Rates

'1'here can be a significant dependence of FET device response on the length of time it takes to accumulate a given total dose, that is on the dose rate of the radiation environment. "1'bus, the total radiation dose alone is not sufficient to specify FET device response to a radiation environment[11]. At low dose rates, the oxide-trapped charge tends to be less and the interface trap density greater than their values at higher dose rates.

Since generally there are changes in device parameters with time during annealing, it is to be expected that during a low dose rate irradiation similar changes will occur in parallel with the damage effects of the radiation[111. 'i'bus, the changes in device parameters measured as a function of total dose will depend on the dose rate at which the radiation is delivered, which is the dose at which it undergoes a given change in parameters. 1 lowever, by characterizing the response with time after the completion of an irradiation, it is possible to predict in a number of circumstances the response as a function of dose rate.

At high dose rates, the response can be dominated by the generation and transport of holes through the oxide, showing large negative threshold voltage shifts immediately after a pulse of radiation; these threshold shifts are reduced with time as the holes transport to the interface [1 1]. With high d ose rates that deliver significant total dose in a short period of time, space-charge effects can reduce the effective electric field in the oxide, reducing the yield of holes which escape recombination and thus, reducing the amount of damage induced by a radiation pulse of given total dose.

There is a reduction in oxide-trapped charge at the lower dose rates and an increase in interface trapped charge [11]. This increase in interface traps as the dose rate decreases is due to the lol]g-term buildup of interface traps. In general, at high dose rates and short times, oxide-trapped charge dominates the device response, whereas at low dose rates or long times, interface trap charge tends to dominate.

Two types of dose rates were tested on the biased chips with gamma irradiation. One chip had the high dose rate of 50 ra(i/see, while the second chip had a lower dose rate of ().1 rad/sec. The results, in contrast to the theoretical expectations, showed no differences between the dose rates. Both chips started degrading at about 5 krads and became nonfunctional at 7 krads. Annealing behavior was about the same with the 50 rad/sec dose rate having a slightly faster and better recovery than the lower dose rate of ().1 rad/sec as observed in the memory tests. It is likely that the (1osc rate was not low enough, and one may need to evaluate characteristics at even lower dose rates of ().()5 to ().()1 rad/sec.

5.4 Degradation

From all the radiation testing, we have observed degradation in the analog and digital portions of the neural network hardware. The analog portion includes the neurons, output buffer, and synapses. The digital portion is mostly the memory in the synapses. Additionally, there are some supporting digital circuits. These include the control and decoding circuits. However, no radiation symptoms were detected in this part of the circuitry. in fact, there is some evidence that the row selection circuit (including a memory latch that held the row address) and the analog output multiplexer were still functioning at a very high radiation dose, while all the other components were severely damaged. We will now discuss the radiation effects seen on the neurons, output buffer, and synapses.

5.4.1 Neuron

Despite the fact that each neuron has a few dozen transistors, they were relatively reliable throughout most of the radiation tests. In the unbiased chips, the neuron sigmoidal curve became steeper at doses greater than 20 krads. This is equivalent to increasing gain of the neuron. For the biased chips, there were no effects at all as seen from neuron () and neuron 1 tests. This might be due to the low doses of radiation (less than 20 krads) or the biasing method. Often the characteristic of the neuron function flattens out, or distorts as the total radiation dose increases due to the damage at the input end in the synapses or the output end in the output buffer. Nevertheless, the neurons usually out-last most of the other components.

S.4.2 Output Buffer

With radiation damage, the output buffer, which is a voltage follower, usually develops a distortion or “kink” in its characteristic function at the high end of its range. The same “kink’> occurs at an input voltage of 7 volts and an output voltage of 6.75 volts for all the different types of radiation. It was observed only in the unbiased chips with a cumulative radiation dose greater than 30 krads. This might also be due to the fact that the biased chips received lower doses of radiation as explained previously for the neurons. All of the neural network’s output must go through this buffer and any degradation would affect the performance of the whole chip. Although the development of the “kink” is not entirely fatal, it may, however, interfere with the “learning” capability of the neural network.

5.4.3 Synapse

Of all the circuits, the behavior of the synapses is most easily and most often disturbed by radiation. They exhibit symptoms of memory failure which contributes to the digital degradation and leakage current which contributes to the analog degradation. The major digital components in the neural network chip are the static memories used within each of the 532 synapses. It was interesting to observe that the memory failures occurred first and most often in the most populated columns of the array (the last 9 columns numbered from 23 to 31) where there were 31 synapses per column. This points out a possible radiation softness in the column-oriented memory architecture. Monotonicity errors also had a column orientation similar to the men-my errors, but this was due to the analog architecture.

The memory circuitry includes a buffer circuit for every column of the memory cells. This buffer will drive the column data line when a memory cell in that column is being written to, and it will buffer data being read out from the column. The unselected memory cells in that column are isolated from the column buffer by a turned "off" (high impedance slate) transmission gate, while the single selected cell is connected to the column buffer through a turned "on" (low impedance state) transmission gate.

During a write operation, the output of the column buffer driver is connected to the output of an inverter in the memory cell through a transmissio n gate in the selected cell. Thus, the column buffer must be strong enough to change the state of the selected memory latch. This is done by forcing the output of a low strength inverter in the memory cell to its opposite state. The column buffer must then be able to source or sink more current than the inverter in the memory latch can output.

After exposure to high doses of radiation, the. transmission gate's high impedance state might not have such a high impedance, causing current to leak into the column data line. Therefore, since the column buffer must source or sink additional current through the leaking transmission gates in addition to the current from the selected memory cell latch, it may not have sufficient strength to flip the bit to the desired state. This condition is worst in the last 9 columns of the synapse array, where there are 31 memory cells in each column.

There arc other similar failure mechanisms that could cause this same behavior, each involving changes in impedance of the transmission gates used in the memory cells. First, the transmission gate could have too high an impedance when it is turned "on" to allow the column buffer to program the memory cell. This condition would add to the difficulty of overcoming the current leaking from turned "off" transmission gates discussed above. Second, leakage in the transmission gates of the unselected cells could overwhelm the output of the selected cell during the read out. This problem would be worst in columns populated with the most memory cells but would not involve the writing circuitry. Any one or a combination of these failure mechanisms can cause the memory errors in the synapses,

The Synapses also demonstrate current leakage as Seen by the left shift not Only in the synapse-neuron curves but in the neuron 1 curve as well. When a synapse is being tested in the synapse-neuron curve test, all the other synapses in the same column are programmed with a DAC weight of zero. This means that all of the n-FET switches in series with the synapse current mirrors (See Figure 1() in Appendix) are turned "off" (high impedance state). With a zero weight, the sign-bit is also zero, and the p-FET current steering mirror is turned "of f." While, the n-FET connecting the p-FET mirror to the summing node of all the n-FET current mirrors is turned "off," the n-FET connecting the summing node of all the n-FET current mirrors to the synapse output is turned "on. " "1'bus, the sign bit being zero favors input of negative current to the neuron from the synapse, which tends to force the neuron output voltage high. If the radiation causes the n-FET switches to leak some current through from the n-FET current mirrors, then each of the synapses programmed to zero will tend to force the neuron in that column high. For higher column number positions, there will be more synapses with a weight of zero leaking current into the current summing input of the neuron in that column. "1'bus, when a synapse in a column is being tested, it would need to have a negative input DAC weight value directly proportional to the number of synapses in that column in order to force the output of the neuron low.

This explains why the synapse-neuron curves progressively shift to the left as the column number increases. It is necessary to program a lower (negative) weight value into the synapse under test to compensate for the leaking synapses and make the total input current into the neuron close to zero where the transition from low to high voltage of the sigmoidal curve occurs.

It is important to note that this column dependent leakage problem is not related to the column dependent memory problem. It is only a coincidence of the specific internal layout of this chip that these problems are both column dependent.

The degradation of the neural network chips whether biased or unbiased for all three types of radiation was mainly due to the radiation effects from the 532 synapses. As seen in the synapse-neuron curves and the neuron 1 curve, the leakage current causes a left shift, forcing the neuron to be turned on with a lower synapse weight value. When the memory errors started occurring, monotonicity errors increased significantly as the synapses were programmed with incorrect values. The sign bit (bit 6) of the memory was shown to have more damage than the other bits. This would cause the synapses to be programmed with a positive value when it should be negative or vice versa, severely changing the shape of the synapse-neuron curves from being sigmoidal. Since the synapses are integrated with the neurons, any degradation would cause the neuron characteristics to change and therefore, alter the performance of the whole chip.

5.5 Power Consumption and Latchup

Power consumption increased with the total dose of radiation. The chip initially used less than 80 mW, but after extreme cumulative doses of radiation, the power consumption would exceed 1,6 Watts in some cases. The chip drew more current as it degraded which might be due to the 532 synapses leaking current. Also, during biased radiation exposure or during the testing procedures, the chip at high doses would latchup and consume more than 1.6 Watts in some cases, making it untestable. Once the chip drew excessive current and the power consumption was greater than 1.6 Watts, it was declared untestable and nonfunctional.

Latchup occurs due to radiation-induced photocurrents generated by high dose rate ionizing radiation exposure [11]. CMOS circuits fabricated on semiconducting substrates are vulnerable to latchup, the state in which a parasitic semiconductor-controlled rectifier turns on. Once latchup is initiated, it fixes circuit nodes in a high-current, low-impedance state, almost "shorting" the Vdd to ground. Activation of this path prevents normal circuit operation and can potentially result in permanent damage similar to that associated with dose-rate hard errors, such as fused interconnect lines or burned-out junctions.

Inherent in the CMOS structure are p-n-p and n-p-n parasitic bipolar transistors. When the "base" region of one device is composed of the same semiconductor material as the "collector" of the other device, these bipolars can form a p-n-p-n type rectifier in parallel with each CMOS inverter.

During normal operation, the p-n-p-n stays in its high impedance, blocking state, and the voltage supplies for the chip pass all their current through the FET structures. If, however, minority carriers are somehow introduced into a bipolar's base region, the p-n-p-n can go into its low impedance state and latch the cell [1 1].

A vertical n-p-n bipolar structure can be turned "on " if any n+ diffusion injects enough electrons into its p-well base. in a positive feedback configuration with the vertical n-p-n is a lateral p-n-p bipolar transistor [11]. Holes from a p+ source/drain diffusion can be injected into the n-substrate base of the p-n-p,

Latchup can be initiated electrically by applying signals to a device, or it can be triggered by radiation events. The semiconduct or-controlled rectifier may be turned "on" by a terminal overvoltage condition caused by a radiation-induced electrical transient or by ionizing radiation characteristic of dose-rate events [11]. Each of these triggering mechanisms injects minority carriers into the base of the parasitic bipolars. ionizing radiation from dose-rate events can inject current across the well/substrate junction. For p-well CMOS, the photocurrents from generated majority carriers, i.e., electrons in the substrate (hole in the well), act as base currents of the parasitic bipolar devices and tend to turn on the rectifier. These effects of ionizing radiation were observed when the biased neural network chips latched-up during exposure at high cumulative doses. However, the chips often became untestable due to latchup after exposure during the application of test signals.

5.6 Annealing

The effects of annealing were significant the first couple of days, especially the first couple of hours, but thereafter remained relatively unchanged for several months. The chip seemed to heal from the radiation damage. For example, the synapse-neuron transfer curves (analog portion) returned to a more functional (sigmoidal) state, and the memory errors (digital portion) decreased. These can be seen in Figures 3.1b, 3.2b, 3.3b, 5.1b, 5.2b, and 5.3b of Chapter 4. However, annealing does not fully restore the chip to the original state even after several months at room temperature. It is interesting to note that the annealing reduced the severity of the radiation damage to the level observed when the chip received a lower dose of radiation. Therefore, this indicates that with annealing the chip can withstand higher total doses of radiation.

Chapter 6

Conclusion and Future Research

Recent technological advances in artificial neural networks have increased the level of interest for critical applications in high dependency environments, such as in intensive care units, operating rooms, emergency rooms, anti recovery rooms of hospitals where reliable performance and longevity is a requirement. Thus, this study allows us to analyze the reliability and robustness of the artificial neural network electronic hardware by studying their performance and degradation characteristics.

The fault tolerant characteristics of artificial neural network electronics is studied by exposing thcm to three types of ionizing radiations: high energy electrons, high energy protons, and gamma radiations. For each experiment, one neural network chip, containing 32 neurons and 532 synapses was biased (with electrical power) while the other chip was unbiased. The results showed that the unbiased chips were able to accumulate an order of a magnitude more radiation dose (1 ()() krads or more) than the biased chips (less than 20 krads) and still remain functional as a neural network.

Both the electron and gamma radiations seemed to cause similar degradation effects on the artificial neural network chips. in contrast, the chips exposed to proton radiation needed higher cumulative doses before showing any degradation, but the damage was more severe.

As the total radiation dose increased, the electronic hardware consisting of analog and digital components showed graceful degradation. For example, the sigmoidal function of the neuron became steeper (analogous to increasing the gain), current leakage from the synapses progressively shifted the synapse-neuron sigmoidal curves to the left, and the digital memory of the synapses and the memory addressing circuits began to gradually fail.

The effects of room temperature annealing were significant the first couple of days following irradiation, especially the first couple of hours, but thereafter remained relatively unchanged for several months. The unbiased chips showed annealing behavior while the biased chips did not show any significant recovery. Both the analog and digital portions of the chip showed some healing. However, annealing does not fully restore the chip to the original state even after several months. Nevertheless, it is interesting to note that the annealing reduced the radiation damage to the effect as if the chip received a lower dose of radiation.

From these radiation experiments, we have a better understanding of the degradation effects on the artificial neural network chips, and the amount of damage (or cumulative radiation dose) needed before the circuitry becomes nonfunctional. Therefore, we can modify certain designs on the neural network chip without using radiation-hardening techniques to enhance the robustness and fault tolerance.

Currently, two of these artificial neural network chips are onboard a *STRV-1b (Space Technology Research Vehicle)* satellite as part of a space radiation experiment.

The chips, flying in a geostationary transfer orbit, pass through the Van Allen belts in space being, exposed to electron and proton radiations. One of the chips is exposed while the other chip is shielded. Both chips remained unbiased most of the time. In a4 month period, the exposed chip has received only 11 krads of total radiation. Thus, in contrast to ground-based radiation study, the space radiation has a lower dose rate (0.875 mrad/see) and the neural network chips have shown only a slight sign of degradation. Also, between exposures there are several hours for the chips to anneal. However, the space experiments gives us the opportunity to (demonstrate the **fault** tolerant characteristics and performance of the artificial neural network chip under realistic application conditions.

in future research studies, using these findings as a guideline, more radiation exposure experiments could be conducted with finer dosages and in a more controlled environment (constant temperature) to carefully study the gradual degradation and the effects of radiation. in addition, degradation could be monitored at both a lower level (i.e., transistor level) and at a higher level (i.e., neural network learning algorithm).

In theory, the response of individual transistors to radiation may be used to elucidate and predict the response of a full integrated circuit to an identical radiation environment. Included with the artificial neural **network** chips are 6 FET devices, which are designed specifically foresting some of the basic effects of radiation. These FET devices represent two technologies (PMOS and NMOS) and three types of devices. The first is a "threshold" dosimeter, which measures the change in the threshold of the FET as a function of total dose. This dosimeter will also provide a means for monitoring the total radiation dose. The second device is a "leakage" dosimeter, which measures the current through a turned-off (gate tied to source) FET by applying a voltage across it. This dosimeter provides useful data on intra-FET (from source to drain within a single transistor) leakage mechanisms. The third device is a "field leakage" dosimeter, which measures the inter-FET (from one transistor to a nearby transistor) leakage current. The results of these FET devices can then be used to provide parameters for modeling transistor characteristics. Using circuit simulation programs such as SPICE, these models can then be utilized to simulate full circuit response with total radiation dose. Therefore, the radiation effects on the neural network chip can be modeled at the transistor level and simulation can be done to not only demonstrate graceful degradation, but also to evaluate and optimize sensitivity to radiation effects.

Learning is the heart of artificial neural networks. By using a learning algorithm, there is a high possibility for the network to adapt to the electronic hardware degradation in the neurons and synapses. For example, the synaptic weights might be able to adjust during training to the increased gain in the neurons and the leakage from the synapses causing a left shift. Also, with a parallel architecture, the neural network will be mm-c fault tolerant with nonfunctional neurons and synapses by compensating thcm with some of the other functional ones. 7'bus, by applying learning to an artificial neural network architecture, there would be higher fault tolerance to the degrading effects of radiation.

References

[1] J. Zurada. *Introduction to Artificial Neural Networks*. West: Minnesota. 1992.

[2] P. K. Simpson. Foundations of Neural Networks. *Artificial Neural Networks, Paradigms, Applications, and Hardware Implemen tations*. Ed: E. Sanchez-Sinencio and C. Lau. IEEE Press: New York. 1992.

[3] P. S. Maclin and J. Dempsey. 1 low to Improve a Neural Network for Early Detection of I lepatic Cancer. *Cancer Letters*. 77(2-3): 95-101, March 15, 1992.

[4] P. S. Maclin and J. Dempsey. Using an Artificial Neural Network to Diagnose I lepatic Masses. *Journal of Medical Systems*. **16(5):**21 5-225, October 1992.

[5] V. Goldberg, A. Manduca, D. L. Ewert, J. J. Gisvold, and J. F. Greenleaf. Improvement in Specificity of Ultrasonography for Diagnosis Of Breast Tumors by Means of Artificial Intelligence. *Medical Physics. 19(6): 1475-1481,* November-December *1992.*

[6] J. A. Scott and E. 1.. Palmer. Neural Network Analysis of Ventilation-Perfusion Lung Stuns. *Radiology.* **186(3):**661 *-664,* March1993.

171 W. G. Baxt. Use of an Artificial Neural Network for the Diagnosis of M yocardial Infarction. *Annals of Internal Medicine. 11* **5(11):** *843-848, December* 1, *1991.*

[8] W. G. Baxt. Analysis of the Clinical Variables Driving Decision in an Artificial Neural Network Trained to Identify the Presence of Myocardial Infarction. *Annals of Emergency Medicine. 21 (12):* 1439-1444, December 1992.

[9] T. Kohonen. *Self-(J/,q[iiliz(lti(~ll and Associative Memory, 3rd Ed.* Springer-Verlag Berlin I leidelberg: New York. 1989.

[1 ()] E. R. Kandel, J. I 1. Schwartz, and T. M. Jesse]]. *Principles of Neural Stir)/ce, 3rd Ed.* Elsevier Science: New York. 1991.

[11] T. P. Ma and P. V. Dressendorfer. *Ionizing Radiation Effects in MOS Devices and Circuits*. Wiley & Sons: New York. 1989.

Appendix

Artificial Neural Network Chip

The artificial neural network chips were specially designed by the Jet Propulsion Laboratory (JPL) for conducting radiation experiments in space onboard a satellite, but were designed and fabricated without any radiation-hardening techniques or any protect ion from radi at ion, A photograph of the actual chip is shown in Figure 8. The chip was fabricated by VI .S1 Technology incorporated (VTI) using a 2-micron N-well CMOS process through MOSIS (MOS Implementation Service). The die size is *7.9* mm *x 9.2 mm* and *is* packaged in a ceramic 64-pin package. The chip requires a single power supply of 8 volts and consumes less than 80 mW of power.

It contains 32 neuron cells and 532 synapse cells arranged in a partially populated (due to power limitations for space flight) 32 row by 32 column array. A block diagram of the layout of the artificial neui al network chip is shown in Figure 1 (Chapter **2). The** neurons lie along a diagonal in the array. The synapses are placed where the row number is less than the column number. However, when the column number is greater than 22, the synapses are placed in every row position except along the diagonal.

Neuron () (bottom left in Figure 1) can only receive input from an external source. Neuron 1 also can receive input externally or from neuron () through the synapse at row 0, column 1. Neurons 2 through 31 receive inputs only from any of the other neurons through the synapses in their respective columns. The first 22 neurons each can receive input from as many synapses as the column number that they are positioned in (i.e., neuron 20, located in column number 20, can receive input from 20 synapses). The last 9 neurons in columns 23 through 31 all have fully connected feedback circuits and thus have 31 synapses connected to the input of each neuron.

The nature of the neural network chip is primarily analog for compactness and low power consumption, with the digital portion only playing a supporting role. The major analog components arc the neurons, synapses, and an output buffer. The major digital components are the static memories for storage of synaptic weight within each of the *532* synapses.

A.1 Neuron

The neurons are nonlinear transimpedance amp] ifiers. The characteristics of an operational amplifier resembles a sigmoidal function and thus is a natural circuit for a neuron. A block diagram of the neuron is shown in the upper right corner of Figure 2 (Chapter 2). Each neuron outputs a voltage which is a sigmoidal function of its input current. A negative input current forces the neuron's output high, and a positive input current forces the output low.

The neuron (based on a former design by JPL [1]) is a typical current summing circuit where the input controls a differential amplifier which controls a current feedback circuit. Figure 9 shows the circuit schematic of the variable-gain, sigmoidal neuron. The current feedback circuit is voltage controlled, and outputs a negative current when given a high input voltage, and outputs a positive current when given a low input voltage. The neuron's current input is connected to the non-inverting input of the differential amplifier. The inverting input is connected to a constant reference voltage of 4 volts, chosen to be the optimum current summing voltage for the synapse outputs and midway between ground and Vdd. If the input current is positive, the input voltage will rise slightly, and the differential amplifier's output voltage will rise sharply and produce a negative feedback current to "cancel out" the positive input current. When this is achieved, (he system is in equilibrium. The operation for a negative input current is correspondingly equivalent.

The differential amplifier and current feedback circuits have high gain so that a few millivolts change on the input will produce a few hundred microampere of feedback current. This allows the input voltage to remain constant within a few millivolts of the reference voltage, as long as the input current is within the operating range (about +/- 500" pA). The consistency of input voltage allows the synapses to reliably output a current extremely close to the current that they are programmed for.

The neuron's output voltage stage is a simple circuit that will Output a voltage which is a sigmoidal function of the input current. The voltage output circuit consists of p-FET and n-FET current mirrors with their drains tied together which mirror the current sources used to feed back current to the neuron input. The drains of these FET devices are also tied to a variable impedance circuit whose function is to control gain or the shape of the neuron's transimpedance function.

If the variable impedance circuit is set to have a high impedance, the transimpedance function will have a very high gain, and the slope in the region where the input current is near zero will be very steep. Thus, it will tend to quickly approach the high asymptote for a negative input current and the negative asymptote for a positive input current. The gain control circuit is controlled by an externally biased current mirror which is common to all the gain **control** circuits in all **of** the neurons. If a high current (-100" pA) is applied to this current mirror, then the gain control circuit connected to the neuron output will have a low impedance, and thus the transimpedance function of the neuron will have a more gradual slope in the operating range (+/- 100 pA). In practice, the gain control is initially adjusted to obtain an output characteristic suitable to the application, and then fixed at this optimum setting.

A voltage to current converter is provided for each row to convert the neuron output voltage to current input for the synapses. This is shown in Figure 10 as part of the synapse circuit schematic. The output voltage of the neuron controls a current into a cascode current mirror which generates current mirror biases for the synapses in that row. Thus, the neuron in each row control s an input current circuit for the synapses in that row. When the voltage output of the neuron is high (above 3 n-FET thresholds), there will be sufficient current input for the synapses in that row to multiply. The higher the neuron output voltage, the more current that will be available for the respective synapses. On the other hand, if the neuron voltage output is low (below 3 n-FET thresholds), there will be virtually no input current to be mirrored, and none of the synapses in that row will output significant current.

A.2 Synapse

The synapse circuit consists of a 7-bit multiplying digital to analog converter (MDAC) and a 7-bit digital memory. This circuit outputs a binary weighted multiple of its analog input current. The input current coining from the neuron in that row is multiplied by the stored digital weight (integer factor between -63 and +63). The 7-bit digital memory, consisting of 7 static latches, provides programmable weight storage and is randomly accessible. The upper left corner of Figure 2 in Chapter 2 shows a block diagram of this synapse. A circuit schematic of this programmable 7-bit synapse is show in Figure 1 (). This design is a modified version of a previous JPL synapse design [2].

Multiplication is accomplished by conditionally scaling the input current by a series of current mirror transistors. For each current mirror, a pass transistor, controlled by one bit of the digital word, conditionally allows current to be placed on a common summation line. The bits in the digital word from the lowest significant bit (1.S13) to the most significant bit (MSB) are connected to 1, 2., 4, 8, 16, and 32 current mirror transistors respectively, so that the input current is scaled by the appropriate amount. The resulting summation current is unipolar. However, a p-FET cascode current mirror, controlled by the seventh bit of the digital word, determines the direction of the output current, such that two quadrant multiplication is accomplished (+/- 64 levels). A negative weight value progran lined into the synapse causes it to have a positive output current (outputs current through its p-FETs) while a positive weight value causes the output to be negative (through the n-FETs to ground).

A synapse may also bc described as a transconductance amplifier, in that it effectively takes a voltage output from a neuron and amplifies it by a conductance (weighted by the digital value stored in the synapse's memory) to produce an output current, In this case we consider the voltage to current converter to be part of the synapse instead of the neuron.

Each synapse’s current output is an input to a neuron in its column. Thus, there is a single current summing node in each column to which all the synapses are directly connected, and that node is also connected to the input of the neuron in (hat column.

A.3 Output Buffer

The output buffer is a wide-range transconductance amplifier configured as a unity-gain follower [3]. The output voltage closely follows its input voltage and provides sufficient driving power to interface the neural network chip to other chips. It can drive a load of approximately 1 ()() kilo-ohms. figure 11 shows the circuit schematic of the output buffer.

The input to the output buffer is the voltage from one of the 31 neurons selected by an analog multiplexer. A special circuit designed into the output multiplexer allows the output buffer to be tested separately by feeding its input with an external] y supplied voltage. The single output buffer and multiplexer design was utilized to simplify the interface between the chip and its external circuitry by reducing the number of separate output signals. The output buffer operated over the entire voltage range of the neuron outputs, except when the voltage approached within a volt of the power supply range of () to 8 volts. Below one volt, the buffer slowed down since the transistors in the circuit began operating in the subthreshold region. The output voltage is generally below the input voltage throughout the operating range, but this error was moderately increased when the input voltage was within one volt of the positive power supply. Since all of the neural network's output must go through this buffer, it is the most critical component (representing a single point failure mechanism for the chip).

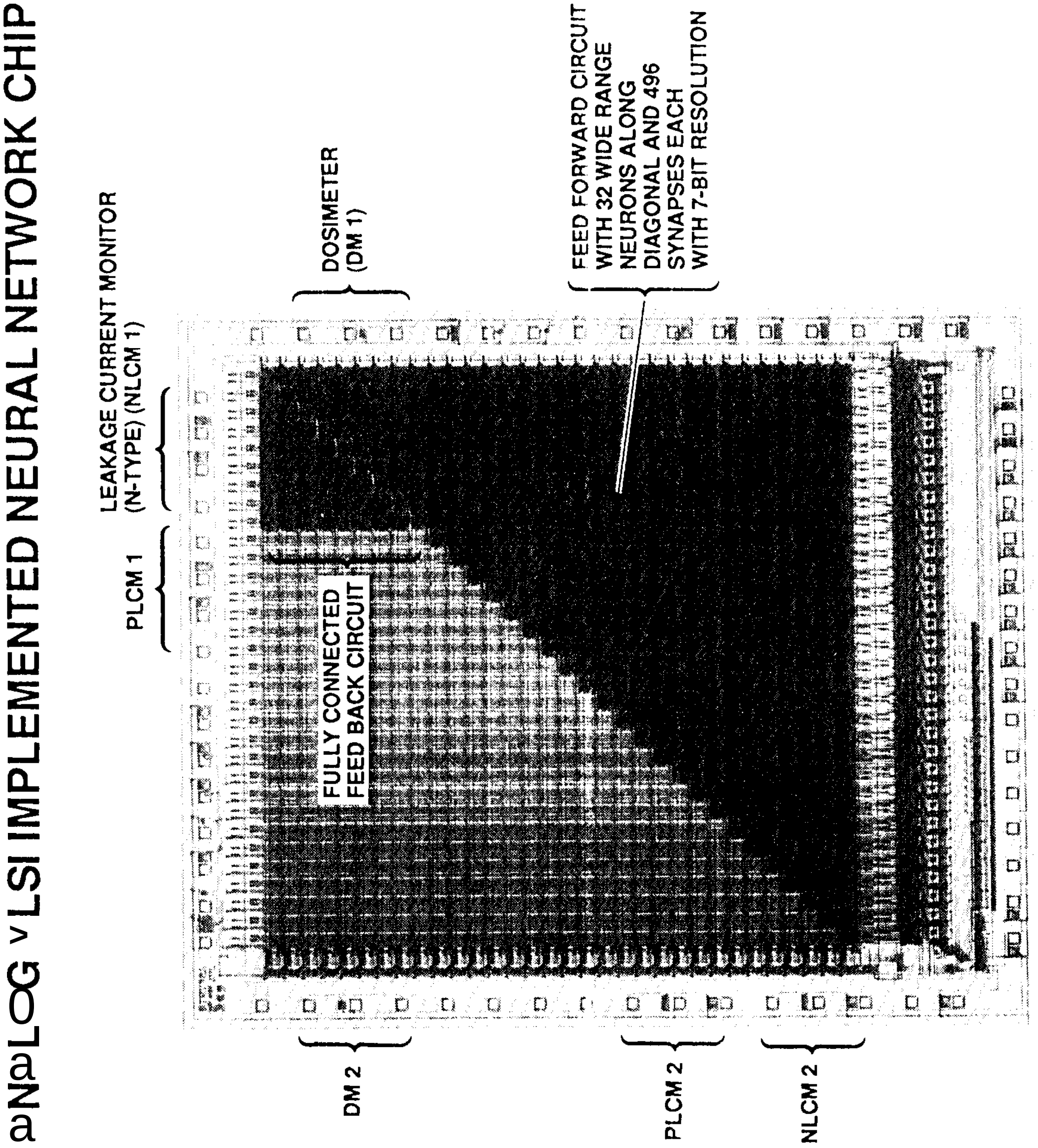
aNaLOG VLSI IMPLEMENTED NEURAL NETWORK CHIP
LEAKAGE CURRENT MONITOR (N-TYPE) (NLCM 1)
PLCM 1
DOSIMETER (DM 1)
FULLY CONNECTED FEED BACK CIRCUIT
FEED FORWARD CIRCUIT WITH 32 WIDE RANGE NEURONS ALONG DIAGONAL AND 496 SYNAPSES EACH WITH 7-BIT RESOLUTION
DM 2
PLCM 2
NLCM 2

Figure 8   Photograph of the Artificial Neural Network Chip

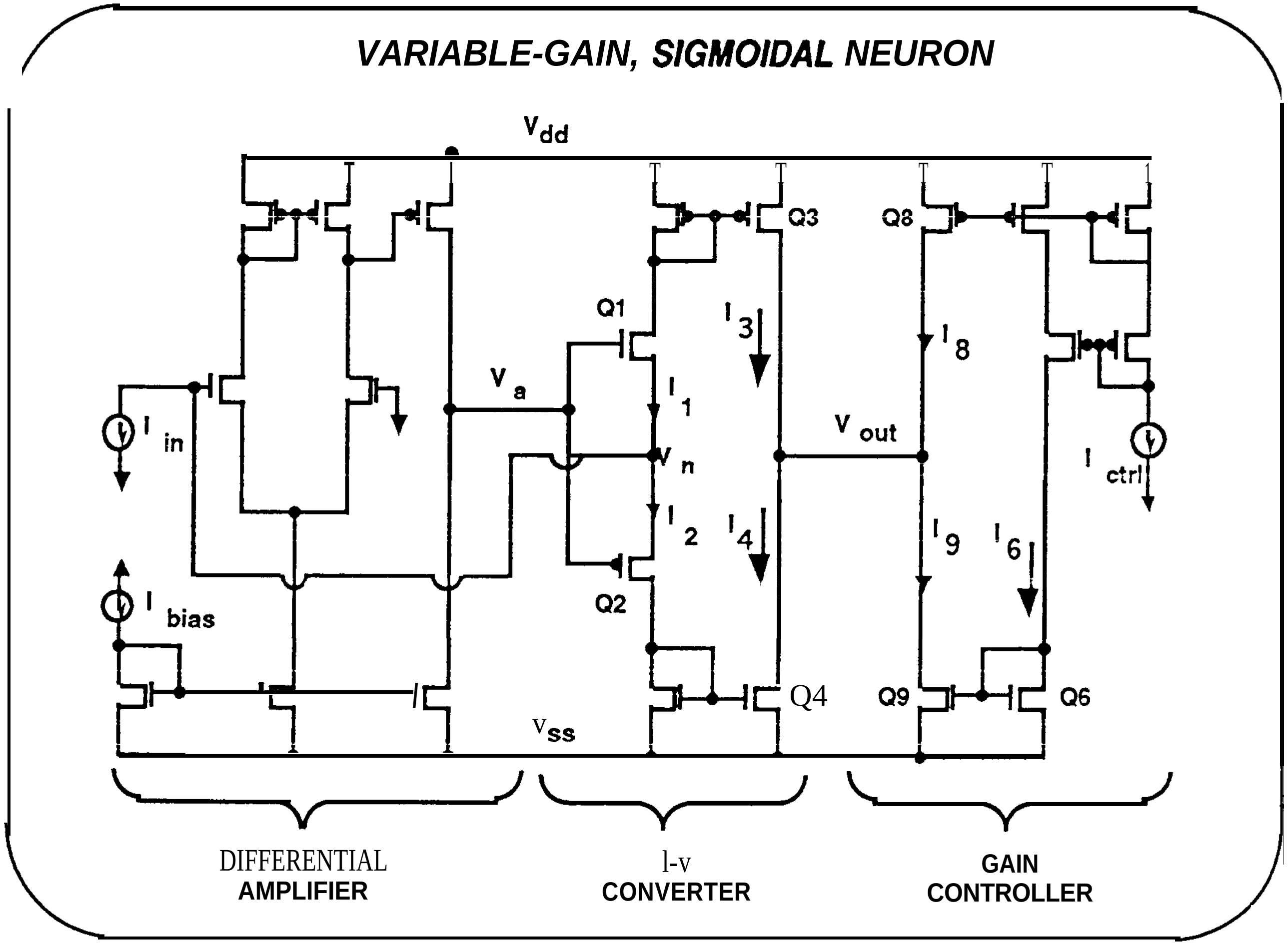

Figure 9 Circuit Schematic of the Variable-Gain, Sigmoidal Neuron

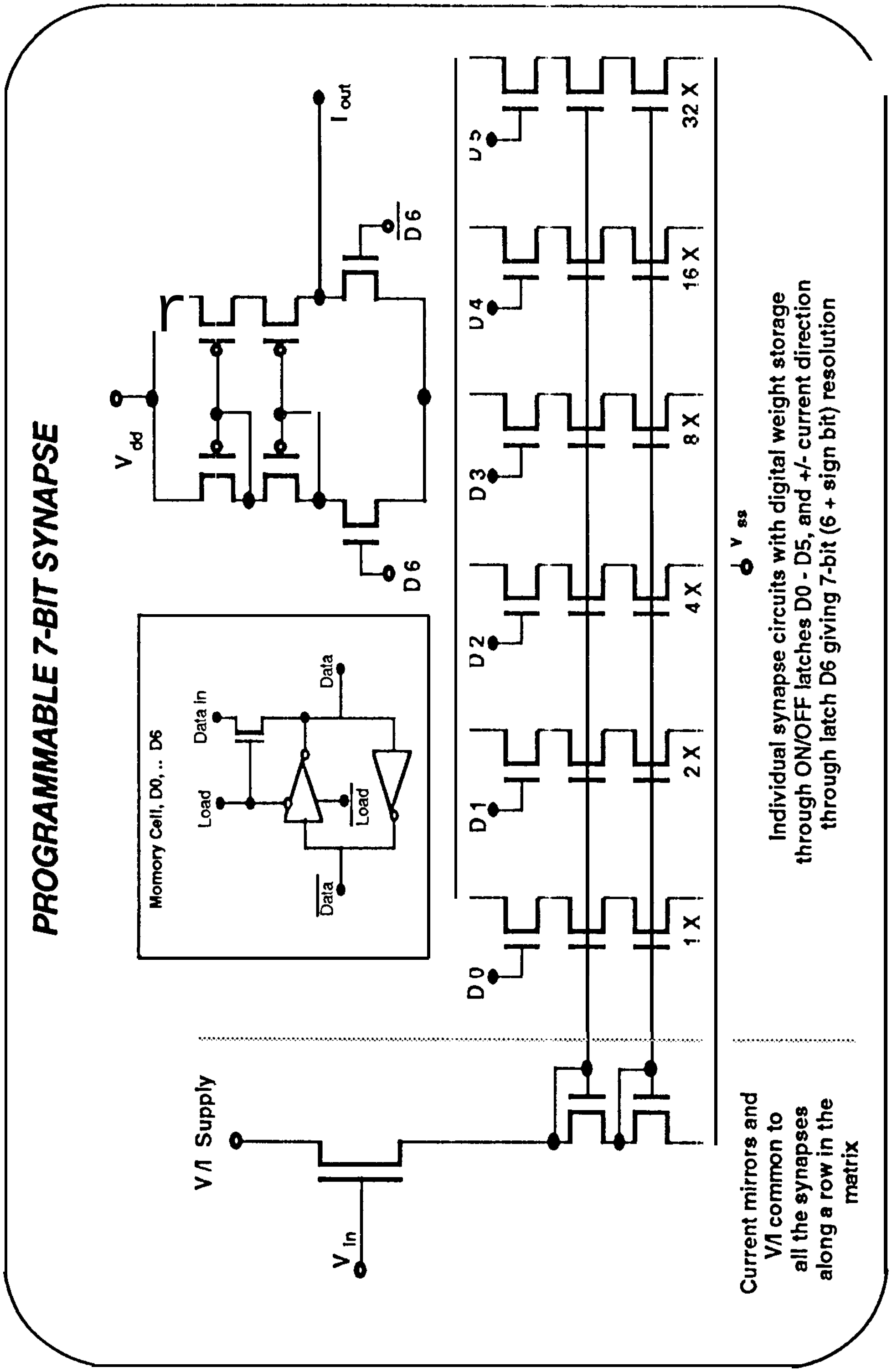
PROGRAMMABLE 7-BIT SYNAPSE
Memory Cell, D0,.. D6
Load
Data in
Data
$\overline{\text{Data}}$
$\overline{\text{Load}}$
V_{dd}
I_{out}
D 6
$\overline{\text{D 6}}$
V/I Supply
V_{in}
D 0
D 1
D 2
D 3
D 4
D 5
1 X
2 X
4 X
8 X
16 X
32 X
V_{ss}
Individual synapse circuits with digital weight storage through ON/OFF latches D0 - D5, and +/- current direction through latch D6 giving 7-bit (6 + sign bit) resolution
Current mirrors and V/I common to all the synapses along a row in the matrix

Figure 10 Circuit Schematic of the Programmable 7-bit Synapse

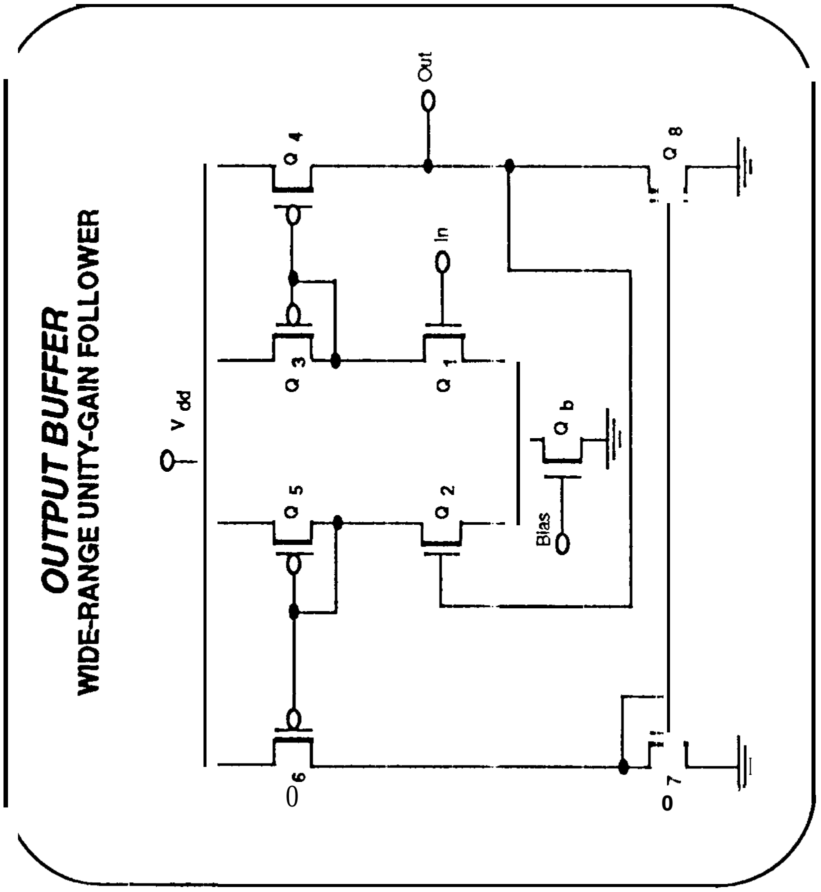

Figure 11 Circuit Schematic of the Output Buffer (Wide-Range Unity-Gain Follower)

A.4 References

[1] T, A. Duong, S. P. Eberhardt, M. D. Tran, T. Daud, and A. P. Thakoor. Learning and Optimization with Cascaded VLSI Neural Network Building-Block Chips. *Proc. IEEE/INNS International Joint Conference on Neural Networks, June 7-11, 1992, Baltimore, MD.* **1**:184-189, *June 1992.*

[2] A. P. Moopenn, 'I'. A. Duong, and A. P. Thakoor. Digital-Analog Hybrid Synapse Chips for Electronic Neural Networks. *Advances in Neural Information Processing Systems (NIPS).* **2**:769-776, 1990.

[3] C. Mead. *Analog VLSI and Neural Systems.* Addison-Wesley: Massachusetts. 1989.